# Partition Models for Variable Selection and Interaction Detection

# Partition Models for Variable Selection and Interaction Detection

A dissertation presented

by

Bo Jiang

to

The Department of Statistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Statistics

Harvard University

Cambridge, Massachusetts

April 2013

**Professor Jun S. Liu** **Bo Jiang**

**Partition Models for Variable Selection and Interaction Detection**

# Abstract

Variable selection methods play important roles in modeling high-dimensional data and are key to data-driven scientific discoveries. In this thesis, we consider the problem of variable selection with interaction detection. Instead of building a predictive model of the response given combinations of predictors, we start by modeling the conditional distribution of predictors given partitions based on responses. We use this inverse modeling perspective as motivation to propose a stepwise procedure for effectively detecting interaction with few assumptions on parametric form. The proposed procedure is able to detect pairwise interactions among $p$ predictors with a computational time of $O(p)$ instead of $O(p^2)$ under moderate conditions. We establish consistency of the proposed procedure in variable selection under a diverging number of predictors and sample size. We demonstrate its excellent empirical performance in comparison with some existing methods through simulation studies as well as real data examples.

Next, we combine the forward and inverse modeling perspectives under the Bayesian framework to detect pleiotropic and epistatic effects in effects in expression quantitative loci (eQTLs) studies. We augment the Bayesian partition model proposed by Zhang et al. (2010) to capture complex dependence structure among gene expression and genetic markers. In particular, we propose a sequential partition prior to model the asymmetric roles played by the response and the predictors, and we develop an efficient dynamic programming algo-

rithm for sampling latent individual partitions. The augmented partition model significantly improves the power in detecting eQTLs compared to previous methods in both simulations and real data examples pertaining to yeast.

Finally, we study the application of Bayesian partition models in the unsupervised learning of transcription factor (TF) families based on protein binding microarray (PBM). The problem of TF subclass identification can be viewed as the clustering of TFs with variable selection on their binding DNA sequences. Our model provides simultaneous identification of TF families and their shared sequence preferences, as well as DNA sequences bound preferentially by individual members of TF families. Our analysis may aid in deciphering *cis* regulatory codes and determinants of protein-DNA binding specificity.

# Contents

# Acknowledgments

This dissertation would not have been possible without the support of many people. I am most grateful to my advisor, Professor Jun S. Liu, for his continuous encouragement and support throughout my years at Harvard. It has been my privilege to have him as a mentor in both academic research and personal life. I thank Professor Joseph Blitzstein and Professor Martha Bulyk for serving on my dissertation committee and providing many invaluable comments on my thesis writing. Professor Martha Bulyk has been a long time collaborator who guided me in the protein binding microarray project. I sincerely appreciate her patience in revising my paper word by word and broadening my scientific knowledge. I would also like to thank National Science Foundation, National Institute of Health and National Human Genome Research Institute for supporting my research with Professor Jun S. Liu and Professor Martha Bulyk through NSF grant 1007762 and NIH/NHGRI R01 HG003985.

Throughout my years at Harvard, it has been my distinct honor to serve as teaching fellows for Professor Joseph Blitzstein, Professor Stephen Blyth and Professor Xiao-Li Meng, who helped me develop various teaching skills and treated me with both food of statistics and food of stomachs. I also thank all the faculty and staff members in the statistics department for providing an incredible learning environment and supportive statistical community.

The long journey of graduate study would be miserable without friends and classmates. I thank Alex Blocker, Ke Deng, Valeria Espinosa, Daniel Fernandez, Simeng Han, Jonathan Hennessy, Ming Hu, Roee Gutman, Joseph Kelly, Yang Li, Cliff Meyer, Nathan Stein, Samuel Wong, Jiexing Wu, Thomas Xiao, Xianchao Xie (a.k.a "XX"), Xiaojin Xu, Yuan Yuan, Anqi

Zhao and Li Zhu for their support, friendship and wisdoms. I also very much enjoyed the companions of undergraduate concentrators, Raj Bhuptani, Jessica Hwang and Keli Liu, who have been my students, TF-mates and teachers.

Last but not least, I owe my greatest gratitude to my parents, Jinhai Jiang and Ru Wang, and my wife, Huayue Zhang, for their love and support. They always encouraged me to pursue what I want and make best efforts to make it happen.

*To my parents.*

# Chapter 1

# Introduction

Recently there has been a significant surge of interest in analytically accurate, numerically robust, and algorithmically efficient variable selection methods, largely due to the tremendous advance in data collection techniques such as those in genetics, internet, and marketing. The importance of discovering truly influential factors from a large pool of possibilities is now widely recognized by both general scientists and quantitative modelers. Motivated by different scientific applications, developments of variable selection techniques encompass various dimensions of statistical modeling including:

- Parametric versus nonparametric models. When a specific form can be derived from solid scientific arguments, parametric models are more accurate in making predictions. In other cases, nonparametric models are more flexible and robust to model mis-specifications, especially in the stage of variable screening and data exploratory analysis. In practice, nonparametric procedures usually involve choosing a discretization (or partition) scheme of continuous data, reflecting a bias-variance trade-off on the resolution of our model assumptions.

- Univariate versus multivariate responses. While it is straightforward to regress each univariate component onto predictors separately, a joint model of multiple responses can reveal relationships among responses and aggregate information from correlated responses to improve the signal strength in selecting important predictor variables.

- Supervised versus unsupervised learning problems. In supervised learning problems, such as classification and regression, the inclusion of redundant variables in the learning procedure may degrade the results. Similarly, in clustering, or unsupervised learning problems, the structure of interest may be best represented using only a few of the feature variables and some form of variable selection prior to, or incorporated into the fitting procedure is advisable.

- Frequentist versus Bayesian approaches. Frequentist methods usually enjoy computational simplicity and asymptotic properties such as consistency. On the other hand, Bayesian framework provides a coherent way to take into account prior knowledge and uncertainties in model selection. Variable selection strategies have benefited from both Frequentist and Bayesian ideas and the interplay of the two standpoints continues to inspire the development of new methods.

In this thesis, we will explore an additional dimension in statistical modeling for variable selection, the forward versus inverse modeling perspective, which leads to the development of partition models for several applications.

## 1.1 An Overview of Variable Selection Methods

Under linear regression models, various regularization methods have been proposed for simultaneously estimating regression coefficients and selecting predictors. Many promising algorithms, such as Lasso (Tibshirani, 1996; Zou, 2006; Friedman et al., 2007), LARS (Efron et al., 2004) and smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001), have been invented. When the number of predictors is extremely large, Fan and Lv (2008) have proposed a sure independence screening (SIS) framework that first independently selects variables based on their correlations with the response and then applies variable selection methods in the second step.

When the relationship between the response $Y$ and predictors $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ is beyond linear, the performance of the variable selection methods based on the linear model assumption can be severely compromised. In his seminal paper on dimension reduction, Li (1991) proposed a semi-parametric index model of the form

$$Y = f(\boldsymbol{\beta}_1^T\mathbf{X}, \boldsymbol{\beta}_2^T\mathbf{X}, \ldots, \boldsymbol{\beta}_q^T\mathbf{X}, \epsilon), \tag{1.1}$$

where $f$ is an unknown link function and $\epsilon$ is a stochastic error independent of $\mathbf{X}$. A sliced inverse regression (SIR) method was developed by Li (1991) to estimate the so-called sufficient dimension reduction (SDR) directions $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q$. Since the estimation of SDR directions does not automatically lead to variable selection, various methods have been developed to perform dimension reduction and variable selection simultaneously in the nonlinear setting. For example, Li (2007) developed sparse SIR (SSIR) algorithm to obtain shrinkage estimates of the SDR directions under $L_1$ norm. Zhong et al. (2012) proposed a stepwise procedure called correlation pursuit (COP) for index models under the SIR framework.

The number of variables to be considered can be even larger with the inclusion of interaction terms between predictors. Consider the following simple example with a regression model for a univariate response variable $Y$ and $p$ independent normally distributed predictor variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$,

$$Y = X_1 X_2 + 0.1\epsilon, \tag{1.2}$$

where $\mathbf{X} \sim \text{MVN}_p(\mathbf{0}, \mathbb{I}_p)$ and $\epsilon \sim \text{N}(0, 1)$. For illustrative purpose, we add a relatively small noise term $0.1\epsilon$ here. Since there are $\binom{p}{2}$ pairwise interactions, fitting regression models with 2-way interactions using variable selection methods, or even the sure independence screening procedure, is challenging when one has a moderate number of predictor variables, say $p = 1000$. Recently, there has been considerable effort in fitting interaction models in the statistical literatures. For example, Bien et al. (2012) developed hierNet, an extension of Lasso to consider interactions in a model if one or both variables are marginally important (referred to as hierarchical interactions by the authors). Li et al. (2012) proposed a sure independence screening procedure based on distance correlation (DC-SIS) that is shown to be capable of detecting important variables when interactions are presented.

## 1.2 Forward versus Inverse Modeling Perspectives

Most of the aforementioned methods are derived from a *forward* modeling perspective, that is, a model for the conditional distribution of $Y$ given $\mathbf{X}$. When predictor variables $\mathbf{X}$ can be treated as random, we obtain a different modeling perspective by "flipping" the roles of $\mathbf{X}$ and $Y$ and putting the response $Y$ behind the (conditioning) bar, which we call an *inverse* model. Indeed, this inverse modeling perspective has been taken by several

researchers and has led to new developments in dimension reduction and variable selection methods. Cook (2007) proposed inverse regression models for dimension reduction, which have deep connections with the SIR method. Simon and Tibshirani (2012) proposed a permutation-based method for testing interactions by exploring the connection between the *forward* logistic model and the *inverse* normal mixture model when the response $Y$ is binary. Another classical method derived from the inverse modeling perspective is the Naïve Bayes classifier for classifications with high-dimensional features. Although Naïve Bayes classifier is limited by its strong independence assumption, it can be generalized by modeling the joint distribution of features. Murphy et al. (2010) proposed a variable selection method using Bayesian information criterion (BIC) for model-based discriminant analysis. Zhang and Liu (2007) proposed a Bayesian method called BEAM to detect epistatic interactions in genome-wide case-control studies, where $Y$ is binary and $\mathbf{X}$ are discrete.

The inverse modeling perspective can also shed lights on interaction detections. Figure 1.1 shows the contour plot for the joint distribution of $Y$ and $X_1$ from the example in (1.2). If we divide the response into five slices and calculate the mean and variance of $X_1$ within each slice (as shown in Figure 1.1), we can see that although $X_1$ is marginally uncorrelated with $Y$ and the conditional means of $X_1$ are the same, the conditional variances of $X_1$ across slices are very different. Instead of screening all the $\binom{p}{2} = O(p^2)$ pairwise interaction terms, we can discover the importance of $X_1$ and $X_2$ by examining the conditional variances of $p$ individual predictors given the sliced response, which only requires a computational complexity of $O(p)$. Motivated by this observation, in this thesis we investigate an inverse modeling approach to attack the problem of interaction detection without imposing model-specific assumptions on the relationship between the response and predictors.

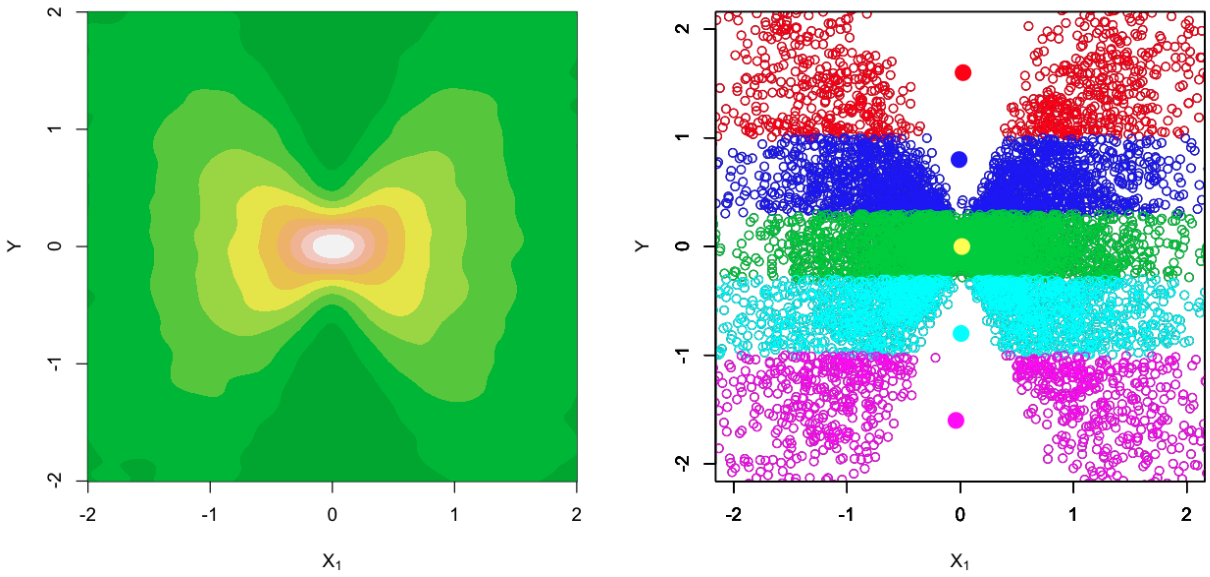


Figure 1.1: Left panel: contour plot for the joint density of $Y$ and $X_1$ in the example (1.2). Right panel: conditional means and variances of $X_1$ given slices of $Y$. Slices are indicated by different colors and round dots mark the conditional means of $X_1$ across slices. The sample conditional variances of $X_1$ within different slices (from top to bottom) are: 2.29, 0.92, 0.41, 0.98 and 2.33.

## 1.3 Integrating Two Perspectives with Partition Model

In Chapter 2, we propose a statistical model for variable selection with interaction detection under the sliced inverse regression framework. A stepwise procedure based on likelihood-ratio test, which we refer to as SIRI, is developed to select relevant predictors from a inverse modeling perspective. Frequentist properties of the proposed variable selection procedure are established and, in particular, its asymptotic behavior under a diverging number of predictors and sample size is investigated. Although the proposed procedure shows promising performance in comparison with existing methods in both simulation studies and real data examples, there are several limitations: (1) the choice of fixed slicing scheme is rather *ad hoc*; (2) stepwise variable selection methods tend to miss important predictor with weak marginal but strong joint effects; (3) there is no straightforward generalization to model

multiple responses and their correlations. These limitations motive us to seek a Bayesian analogue of the slice inverse regression framework.

Under a Bayesian framework, we generalize the concept of slice by introducing a latent variable $T$ that we call *individual type*. Conditioning on individual type $T$, we first decouple the forward and inverse perspectives by temporarily releasing both $X$ and $Y$ out of the (conditioning) bar and modeling them separately. Then, we combine both models to draw posterior inferences on $T$ given its prior distribution. The question that remains is how best to model hidden individual type $T$. The Dirichlet process prior, a common choice in unsupervised Bayesian clustering, is flexible but not suitable for modeling individual types in regression problems. Since our objective here is to extract information from predictors $X$ that can be used to explain variation in the response $Y$, the Dirichlet process prior gives $X$ and $Y$ too much "freedom" and does not respect the asymmetric roles played by $X$ and $Y$. To accommodate the objective of supervised learning, in Chapter 3, we propose a sequential partition prior for modeling individual type $T$ and an efficient sampling algorithm based on dynamic programming.

## 1.4 Applications in Genetics and Gene Transcriptional Regulations

Developments of statistical models and variable selection methods in this thesis are motivated by genetic studies in expression quantitative trait loci (eQTL) and biological research in gene transcriptional regulation.

Expression quantitative trait loci (eQTLs) are genomic loci that regulate expression levels of genes. By assaying gene expression and genetic variation simultaneously on a genome-

wide basis, scientists wish to discover groups of genomic loci (among millions) that can affect the expression of a subset of genes (among thousands). The problem can be viewed as a multivariate regression with variable selection on both responses (gene expression) and predictors (genetic markers), including also multi-way interactions among correlated predictors (called *epistatic effects* of genetic markers). Motivated by BEAM model (Zhang and Liu, 2007) for genome-wide case-control studies, Zhang et al. (2010) proposed a Bayesian partition model for eQTL studies with continuous responses (expression levels of genes) and discrete predictors (genotypes of genetic markers).

Transcription factors (TFs) regulate the expression of their target genes through interactions with specific DNA binding sites in the genome. In our effort to generalize the partition model to identify key regulatory TFs in embryonic stem cells, when both the response (gene expression) and predictors (TF binding specificities) are continuous, we found interesting links with sliced inverse regression (SIR) and correlation pursuit (COP) that finally lead to the development of a procedure for detecting interactions from an inverse modeling perspective, which we named SIRI in Chapter 2. The proposed procedure can be viewed as a Frequentist analogue of the Bayesian partition model (Zhang et al., 2010) for continuous response and predictor variables.

TFs with similar structures can be divided into subclasses, with more closely related proteins exhibiting more similar DNA binding preferences. Numerous studies have employed universal protein binding microarray (PBM) (Berger et al., 2006) technology to determine the *in vitro* binding specificities of hundreds of TFs for all possible 8-bp DNA sequences (8-mers). Identification of TF subclasses based on PBM 8-mer data can be viewed as an unsupervised learning (clustering) problem with variable selections. In Chapter 4, we generalize the Bayesian partition model to provide simultaneous identification of TF subclasses

and their shared sequence preferences, and also of 8-mers bound preferentially by individual members of TF subclasses. Such results may aid in deciphering *cis* regulatory codes and determinants of protein-DNA binding specificity.

The development of SIRI and its application in transcriptional regulation also provide new research ideas on improving the partition model for eQTL studies. Inspired by a dynamic slicing scheme designed for SIRI, we propose a sequential partition prior, which accounts for the asymmetric roles played by response and predictors in eQTL problems, and develop an efficient dynamic programming algorithm for sampling latent individual partitions in Chapter 3. We further augment the Bayesian partition model to capture complex dependence structure among gene expression and genetic markers, especially negative co-expression between genes and linkage disequilibrium (LD) between genetic markers. Through simulation studies and a real data example in yeast, we demonstrate how the augmented Bayesian partition model achieved significantly improved power in detecting eQTLs compared to the original model proposed by Zhang et al. (2010) and other traditional methods.

## 1.5 Outline

An outline for the remainder of this dissertation is as follows. Chapter 2 gives a detailed account of the SIRI method for variable selection and interaction detection with comparisons and connections to previous methods. Chapter 3 elaborates the Bayesian partition model and its extension for identifying pleiotropic and epistasis eQTLs. Chaper 4 studies the application of partition model in unsupervised learning of TF subclasses based on PBM $k$-mer data. Chapter 5 summarizes, pointing out some directions for future research.

# Chapter 2

# Sliced Inverse Regression with Interaction Detection

Let $Y \in \mathbb{R}$ be a univariate response variable and $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T \in \mathbb{R}^p$ be a vector of $p$ continuous predictor variables with covariance matrix $\Sigma_p = \mathrm{Cov}(\mathbf{X})$. $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are independent observations of $(\mathbf{X}, Y)$. For discrete response, we can naturally group $\{y_i\}_{i=1}^n$ into a finite number of classes. For continuous response, the range of $\{y_i\}_{i=1}^n$ can be divided into $H$ disjoint intervals, referred to as slices, which are denoted as $S_1, S_2, \ldots, S_H$. We define a random variable $S(Y)$ as the slice membership of response $Y$, that is, $S(Y) = h$ if $Y \in S_h$. For a fixed slicing scheme, we denote $n_h = |S_h| = s_h n$ where $\sum_{h=1}^H s_h = 1$.

The working model of the sliced inverse regression (SIR) method is given by (1.1). After standardization, $\mathbf{Z} = \Sigma_p^{-\frac{1}{2}} [\mathbf{X} - \mathbb{E}(\mathbf{X})]$, we can rewrite model (1.1) as

$$Y = f\left(\boldsymbol{\eta}_1^T \mathbf{Z}, \boldsymbol{\eta}_2^T \mathbf{Z}, \ldots, \boldsymbol{\eta}_q^T \mathbf{Z}, \epsilon\right) \text{ with } \boldsymbol{\eta}_i = \Sigma_p^{\frac{1}{2}} \boldsymbol{\beta}_i. \tag{2.1}$$

where $f$ is an unknown link function and $\epsilon$ is a stochastic error independent of $\mathbf{X}$. The SIR al-

gorithm was motivated by the following observation: under model (2.1), if $\mathbb{E}\left(\mathbf{b}^T\mathbf{Z}|\boldsymbol{\eta}_1^T\mathbf{Z},\ldots,\boldsymbol{\eta}_q^T\mathbf{Z}\right)$ is linear in $\boldsymbol{\eta}_1^T\mathbf{Z},\ldots,\boldsymbol{\eta}_q^T\mathbf{Z}$ for any vector $\mathbf{b}\in\mathbb{R}^p$, then $\mathbb{E}\left(\mathbf{Z}|Y\right)$ is contained in the linear subspace spanned by $\boldsymbol{\eta}_1,\ldots,\boldsymbol{\eta}_q$ (Li, 1991). So if the first $q$ largest eigenvalues of $\mathrm{Cov}\left(\mathbb{E}\left(\mathbf{Z}|Y\right)\right)$ are all positive, SDR directions can be obtained by the corresponding eigenvectors.

Eigenvalues of $\mathrm{Cov}\left(\mathbb{E}\left(\mathbf{Z}|Y\right)\right)$ also connect SIR with multiple linear regression (MLR). In MLR, the correlation squared, $R^2$, can be defined as

$$R^2=\max_{\mathbf{b}\in\mathbb{R}^p}\left[\mathrm{Corr}\left(Y,\mathbf{b}^T\mathbf{Z}\right)\right]^2,$$

while in SIR, the largest eigenvalue of $\mathrm{Cov}\left(\mathbb{E}\left(\mathbf{Z}|Y\right)\right)$, called the first *profile-*$R^2$, can be defined as

$$\lambda_1\left(\mathrm{Cov}\left(\mathbb{E}\left(\mathbf{Z}|Y\right)\right)\right)=\max_{\mathbf{b}\in\mathbb{R}^p}\max_{T}\left[\mathrm{Corr}\left(T(Y),\mathbf{b}^T\mathbf{Z}\right)\right]^2,$$

where the maximum is taken over all bounded transformations $T(\cdot)$ and vectors $b\in\mathbb{R}^p$ (Chen and Li, 1998). We can further define the $k$th profile-$R^2$, $\lambda_k$ $(2\le k\le q)$, as the $k$th largest eigenvalue of $\mathrm{Cov}\left(\mathbb{E}\left(\mathbf{Z}|Y\right)\right)$ by restricting the vector $\mathbf{b}$ to be orthogonal to eigenvectors of the first $(k-1)$ profile-$R^2$. Building upon this connection with MLR, Zhong et al. (2012) proposed the correlation pursuit (COP) method for variable selection motivated by the F-test in stepwise regression. The COP statistic is defined as

$$\mathrm{COP}_k^{d+1}=n\frac{\widehat{\lambda}_k^{d+1}-\widehat{\lambda}_k^{d}}{1-\widehat{\lambda}_k^{d+1}},k=1,2,\ldots,q,\ \text{and}\ \mathrm{COP}_{1:q}^{d+1}=\sum_{k=1}^{q}\mathrm{COP}_k^{d+1},$$

where $\widehat{\lambda}_k^{d}$ and $\widehat{\lambda}_k^{d+1}$ are the $k$th profile-$R^2$ estimated from the current set of selected predictors of dimension $d$ and current selected predictors plus an additional predictor to be considered, respectively. A stepwise procedure based on the COP statistic was developed in Zhong et al.

(2012) for simultaneous dimension reduction and variable selection.

## 2.1 From SIR to SIRI: An Inverse Model for Variable Selection and Interaction Detection

To take an inverse perspective on SIR, we start with a seemingly different model. We assume that the distribution of predictors given the sliced response is multivariate normal:

$$\mathbf{X}|Y \in S_h \sim \mathrm{MVN}\left(\mu_h, \Sigma\right), 1 \leq h \leq H, \tag{2.2}$$

where $\mu_h - \mu \in \mathbb{V}^q$ belongs to a $q$-dimensional affine space, $\mathbb{V}^q$ is a $q$-dimensional subspace ($q < p$) and $\mu \in \mathbb{R}^p$. Alternatively, we can write $\mu_h = \mu + \Gamma\gamma_h$, where $\gamma_h \in \mathbb{R}^q$ and $\Gamma$ is a $p$ by $q$ matrix whose columns form a basis of the subspace $\mathbb{V}^q$. Although this representation is only unique up to a orthonormal transformation on the bases $\Gamma$, the subspace $\mathbb{V}^q$ is unique and identifiable. The following proposition proved by Szretter and Yohai (2009) links the inverse model (2.2) with SIR.

**Proposition 1.** *The maximum likelihood estimate (MLE) of the bases of subspace $\mathbb{V}^q$ in model (2.2) coincides with the SDR directions estimated from the SIR algorithm up to a orthonormal transformation.*

According to Proposition 1, we could have derived the SIR algorithm from an inverse model. Next, we propose to select variables via a hypothesis testing framework based on this model.

### 2.1.1 Likelihood-Ratio Tests for Selecting Variables with Marginal Effects

Here we provide a view of the COP method for selecting variables with marginal effects from an inverse modeling perspective. This will lay the ground work for the augmented model to select variables with interaction effects in the next section.

For the purpose of variable selection, we partition predictors into two subsets: a set of relevant predictors indexed by $\mathcal{A}$ and a set of redundant predictors indexed by $\mathcal{A}^c$, and assume the following model:

$$\begin{aligned} \mathbf{X}_{\mathcal{A}}|Y \in S_h &\sim \text{MVN}\left(\mu_h \in \mu + \mathbb{V}^q, \Sigma\right), \\ \mathbf{X}_{\mathcal{A}^c}|\mathbf{X}_{\mathcal{A}}, Y \in S_h &\sim \text{MVN}\left(\alpha + \boldsymbol{\beta}^T \mathbf{X}_{\mathcal{A}}, \Sigma_0\right). \end{aligned} \tag{2.3}$$

That is, we assume that the conditional distribution of relevant predictors follows the inverse model (2.2) of SIR and has a common covariance matrix in different slices. Given the current set of selected predictors indexed by $\mathcal{C}$ with dimension $d$ and another predictor indexed by $j \notin \mathcal{C}$, we propose the following hypotheses:

$$H_0 : \mathcal{A} = \mathcal{C} \text{ v.s. } H_1 : \mathcal{A} = \mathcal{C} \cup \{j\}.$$

Let $\Theta_0$ and $\Theta_1$ denote the parameter space under $H_0$ and $H_1$, respectively. The likelihood-ratio test statistic can be written as

$$L_{j|\mathcal{C}} = \frac{\max_{\theta_1 \in \Theta_1} \left(P_{\theta_1}\left(\mathbf{X}|Y\right)\right)}{\max_{\theta_0 \in \Theta_0} \left(P_{\theta_0}\left(\mathbf{X}|Y\right)\right)} = \frac{P_{\widehat{\theta}_1}\left(\mathbf{X}_{[\mathcal{C}\cup\{j\}]^c}|X_j, \mathbf{X}_{\mathcal{C}}, Y\right) P_{\widehat{\theta}_1}\left(X_j|\mathbf{X}_{\mathcal{C}}, Y\right)}{P_{\widehat{\theta}_0}\left(\mathbf{X}_{[\mathcal{C}\cup\{j\}]^c}|X_j, \mathbf{X}_{\mathcal{C}}, Y\right) P_{\widehat{\theta}_0}\left(X_j|\mathbf{X}_{\mathcal{C}}, Y\right)} = \frac{P_{\widehat{\theta}_0}\left(X_j|\mathbf{X}_{\mathcal{C}}, Y\right)}{P_{\widehat{\theta}_0}\left(X_j|\mathbf{X}_{\mathcal{C}}, Y\right)},$$

where $\widehat{\theta}_0 = \text{argmax}_{\theta_0 \in \Theta_0} \left( P_{\theta_0} \left( \mathbf{X} | Y \right) \right)$, $\widehat{\theta}_1 = \text{argmax}_{\theta_1 \in \Theta_1} \left( P_{\theta_1} \left( \mathbf{X} | Y \right) \right)$, and the last equality follows from $P_{\widehat{\theta}_1} \left( \mathbf{X}_{[\mathcal{C} \cup \{j\}]^c} | X_j, \mathbf{X}_{\mathcal{C}}, Y \right) = P_{\widehat{\theta}_0} \left( \mathbf{X}_{[\mathcal{C} \cup \{j\}]^c} | X_j, \mathbf{X}_{\mathcal{C}}, Y \right)$ according to model (2.3). The scaled log-likelihood-ratio test statistic is given by

$$\widehat{D}_{j|\mathcal{C}} = \frac{2}{n} \log \left( L_{j|\mathcal{C}} \right) = \sum_{k=1}^{q} \log \left( 1 + \frac{\widehat{\lambda}_k^{d+1} - \widehat{\lambda}_k^{d}}{1 - \widehat{\lambda}_k^{d+1}} \right), \tag{2.4}$$

where $\widehat{\lambda}_k^d$ and $\widehat{\lambda}_k^{d+1}$ are estimates of the $k$th profile-$R^2$ based on $\mathbf{x}_{\mathcal{C}}$ and $\mathbf{x}_{\mathcal{C} \cup \{j\}}$, respectively. Under the null hypothesis, $\frac{\widehat{\lambda}_k^{d+1} - \widehat{\lambda}_k^{d}}{1 - \widehat{\lambda}_k^{d+1}} \xrightarrow{P} 0$. Since $\log(1+t) \approx t$ when $t$ is small, we have

$$\left( n \widehat{D}_{j|\mathcal{C}} \right) \xrightarrow{P} \text{COP}_{1:q}^{d+1} = \sum_{k=1}^{q} \text{COP}_k^{d+1} \xrightarrow{D} \chi_q^2,$$

as $n \to \infty$. Coincidentally, we re-discovered the COP statistic of Zhong et al. (2012) from an inverse model. For all the predictors indexed by $j \in \mathcal{C}^c$, we can also obtain the asymptotic joint distribution of $\left( n \widehat{D}_{j|\mathcal{C}} \right)$ under the null hypothesis:

$$\left( n \widehat{D}_{j|\mathcal{C}} \right)_{j \in \mathcal{C}^c} \xrightarrow{D} \left( \sum_{k=1}^{K} z_{kj}^2 \right)_{j \in \mathcal{C}^c} \tag{2.5}$$

where $\mathbf{z}_k = \left( z_{kj} \right)_{j \in \mathcal{C}^c} \sim \text{MVN} \left( \mathbf{0}, \left[ \text{Corr} \left( X_i, X_j | \mathbf{X}_{\mathcal{C}} \right) \right]_{i,j \in \mathcal{C}^c} \right)$ and $\mathbf{z}_k$'s are independent.

Furthermore, as $n \to \infty$,

$$\widehat{D}_{j|\mathcal{C}} \xrightarrow{a.s.} D_{j|\mathcal{C}} = \log \left( 1 + \frac{\text{Var}(M_j) - \text{Cov} \left( M_j, \mathbf{X}_{\mathcal{C}} \right) \left[ \text{Cov} \left( \mathbf{X}_{\mathcal{C}} \right) \right]^{-1} \text{Cov} \left( M_j, \mathbf{X}_{\mathcal{C}} \right)^T}{V_j} \right)$$

where $M_j = \mathbb{E} \left( X_j | \mathbf{X}_{\mathcal{C}}, S(Y) \right)$, $V_j = \text{Var} \left( X_j | \mathbf{X}_{\mathcal{C}}, S(Y) \right)$ and $S(Y) = h$ when $Y \in S_h$ ($1 \leq h \leq H$). Note that $V_j$ does not depend on $Y$ under the assumption in model (2.3). By the

Cauchy-Schwarz inequality and normality assumption,

$$D_{j|\mathcal{C}} = 0 \text{ iff } \mathbb{E}\left(X_j|\mathbf{X}_{\mathcal{C}}, Y \in S_h\right) = \mathbb{E}\left(X_j|\mathbf{X}_{\mathcal{C}}\right), 1 \leq h \leq H.$$

That is, the test statistic $\widehat{D}_{j|\mathcal{C}}$ almost surely converges to zero if the conditional mean of $X_j$ is independent of slice membership $S(Y)$. Detailed proofs on properties of likelihood-ratio test statistic are delegated to Appendix A.1.

Given thresholds $\nu_a > \nu_d$ and the current set of selected predictors indexed by $\mathcal{C}$, we can select relevant variables by iterating the following steps until no new addition or deletion occurs:

- Addition step: find $j_a$ such that $\widehat{D}_{j_a|\mathcal{C}} = \max_{j\in\mathcal{C}^c} \widehat{D}_{j|\mathcal{C}}$; if $\widehat{D}_{j_a|\mathcal{C}} > \nu_a$, let $\mathcal{C} = \mathcal{C} + \{j_a\}$.
- Deletion step: find $j_d$ such that $\widehat{D}_{j_d|\mathcal{C}-\{j_d\}} = \min_{j\in\mathcal{C}} \widehat{D}_{j|\mathcal{C}-\{j\}}$; if $\widehat{D}_{j_d|\mathcal{C}-\{j_d\}} < \nu_d$, let $\mathcal{C} = \mathcal{C} - \{j_d\}$.

Under model (2.3), for relevant predictors indexed by $j \in \mathcal{A}$, we have

$$X_j|\mathbf{X}_{\mathcal{A}-\{j\}}, Y \in S_h \sim \mathrm{N}\left(\alpha_j^{(h)} + \boldsymbol{\beta}_j^T \mathbf{X}_{\mathcal{A}-\{j\}}, \sigma_j^2\right), 1 \leq h \leq H. \tag{2.6}$$

Let $\alpha_j(Y) = \sum_{h=1}^{H} \alpha_j^{(h)} \mathbb{I}\left(Y \in S_h\right)$. We introduce the following concept to study the marginal effect of relevant predictors.

**Definition 1** (Marginally Detectable)**.** *We say a predictor indexed by $j$ is* marginally detectable *if there exist constants $\kappa \geq 0$ and $\xi > 0$ such that*

$$Var\left(\alpha_j(Y)\right) \geq \xi n^{-\kappa}, \tag{2.7}$$

*for all $n$.*

Under the following conditions, the stepwise procedure proposed above is consistent by choosing the thresholds appropriately.

**Condition 1.** *There exist $0 < \tau_{min} < \tau_{max} < \infty$ such that*

$$\tau_{min} \leq \lambda_{min}\left(\mathrm{Cov}\left(\mathbf{X}|Y \in S_h\right)\right) < \lambda_{max}\left(\mathrm{Cov}\left(\mathbf{X}|Y \in S_h\right)\right) \leq \tau_{max},$$

*and*

$$\lambda_{max}\left(\mathrm{Cov}\left(\mathbf{X}\right)\right) \leq \tau_{max},$$

*where $\lambda_{min}(\cdot)$ and $\lambda_{max}(\cdot)$ denote the smallest and largest eigenvalue of a positive definite matrix.*

**Condition 2.** *$p = O(n^{\rho})$ as $n \to \infty$ with $\rho > 0$ and $2\rho + 2\kappa < 1$, where $\kappa$ is the same constant as in (2.7).*

The following theorem is proved in Appendix A.2.

**Theorem 1.** *Under Condition 1 and Condition 2, if all the relevant predictors indexed by $\mathcal{A}$ in model (2.3) are marginally detectable with constant $\kappa$, then there exists constant $c > 0$ such that*

$$\Pr\left(\min_{\mathcal{C}:\mathcal{C}^c \cap \mathcal{A} \neq \emptyset} \max_{j \in \mathcal{C}^c} \widehat{D}_{j|\mathcal{C}} \geq cn^{-\kappa}\right) \to 1, \text{ and } \Pr\left(\max_{\mathcal{C}:\mathcal{C}^c \cap \mathcal{A} = \emptyset} \max_{j \in \mathcal{C}^c} \widehat{D}_{j|\mathcal{C}} < \frac{c}{2} n^{-\kappa}\right) \to 1,$$

*as $n \to \infty$.*

Thus, if we choose the threshold $\nu_a = cn^{-\kappa}$ and $\nu_d = (c/2)n^{-\kappa}$ with $c$ and $\kappa$ defined above, then the addition step will not stop selecting variables until all the relevant predictors have

been included, and once all the relevant predictors have been included, all the redundant variables will be removed from the selected variables.

### 2.1.2 An Augmented Model with Interaction Detection

Let us revisit the simple example in (1.2) with the stepwise procedure based on the likelihood-ratio test statistic proposed in the previous section. In this example, $\mathbb{E}(X_1|Y \in S_h) = \mathbb{E}(X_2|Y \in S_h) = 0$ for $1 \leq h \leq H$. In the first addition step with $\mathcal{C} = \emptyset$, the stepwise procedure fails to capture either $X_1$ or $X_2$ since $D_{1|\mathcal{C}=\emptyset} = D_{2|\mathcal{C}=\emptyset} = 0$. In order to detect predictors with interactions, such as $X_1$ and $X_2$ in this example, we augment model (2.3) to a more general form:

$$\begin{aligned} \mathbf{X}_{\mathcal{A}}|Y \in S_h &\sim \text{MVN}(\mu_h, \Sigma_h), \\ \mathbf{X}_{\mathcal{A}^c}|\mathbf{X}_{\mathcal{A}}, Y \in S_h &\sim \text{MVN}\left(\alpha + \boldsymbol{\beta}^T\mathbf{X}_{\mathcal{A}}, \Sigma_0\right), \end{aligned} \tag{2.8}$$

which differs from model (2.3) in its allowing for slice-dependent means and covariance matrices for relevant predictors.

Under model (2.8), a predictor indexed by $j \in \mathcal{A}$ is *conditionally irrelevant* if the conditional distribution of $X_j$ given $X_{\mathcal{A}-\{j\}}$ and $S(Y)$ does not depend on slice $S(Y)$. If there exists a conditionally irrelevant predictor indexed by $j \in \mathcal{A}$, then we can always redefine the index set of relevant predictors to be $\mathcal{A} - \{j\}$ in model (2.8). To guarantee identifiability, variables indexed by $\mathcal{A}$ in model (2.8) have to be *minimally relevant*, that is, $\mathcal{A}$ does not contain any conditionally irrelevant predictor. For example, if the joint distribution of $(X_1, X_2)$ depends on $S(Y)$ and all the other variables are conditionally independent of $S(Y)$ given $X_1$, then $X_3$ is conditionally irrelevant given $(X_1, X_2)$ and so $\{X_1, X_2, X_3\}$ is relevant but not

minimally relevant. In this example, $\{X_1, X_2\}$ is minimally relevant if both the conditional distribution of $X_1$ given $X_2$ and the conditional distribution of $X_2$ given $X_1$ depend on $S(Y)$. In Appendix A.3, we prove the following proposition:

**Proposition 2.** *The set of minimally relevant predictors indexed by $\mathcal{A}$ under model (2.8) is unique given Condition* 1*.*

By following the same hypothesis testing framework for variable selection in the previous section, we can derive the scaled log-likelihood-ratio test statistic under the augmented model (2.8):

$$\widehat{D}^*_{j|\mathcal{C}} = \log \widehat{\sigma}^2_j - \sum_{h=1}^{H} \frac{n_h}{n} \log \left[\widehat{\sigma}^{(h)}_j\right]^2, \tag{2.9}$$

where $\mathcal{C}$ indexes currently selected predictors and $j \in \mathcal{C}^c$, $\left[\widehat{\sigma}^{(h)}_j\right]^2$ is the estimated variance by regressing $X_j$ on $\mathbf{X}_\mathcal{C}$ in slice $S_h$, and $\widehat{\sigma}^2_j$ is the estimated variance by regressing $X_j$ on $\mathbf{X}_\mathcal{C}$ using all the observations. Under the assumption that $\mathcal{A} \subset \mathcal{C}$ with $|\mathcal{C}| = d$, we can derive the exact and asymptotic distribution of $\left(n\widehat{D}^*_{j|\mathcal{C}}\right)$:

$$n\widehat{D}^*_{j|\mathcal{C}} \sim n \log\left(1 + \frac{Q_0}{\sum_{h=1}^{H} Q_h}\right) - \sum_{h=1}^{H} \frac{n_h}{n} \log\left(\frac{Q_h/n_h}{\sum_{h=1}^{H} Q_h/n}\right) \xrightarrow{D} \chi^2_{(H-1)(d+2)},$$

where $Q_0 \sim \chi^2_{(H-1)(d+1)}$ and $Q_h \sim \chi^2_{n_h-(d+1)}$ $(1 \le h \le H)$ are mutually independent according to *Cochran's theorem*. For all the predictors indexed by $j \in \mathcal{C}^c$ given predictors indexed by $\mathcal{C}$, we can also obtain the asymptotic joint distribution of $\left(n\widehat{D}^*_{j|\mathcal{C}}\right)$ under the assumption that $\mathcal{A} \subset \mathcal{C}$:

$$\left(n\widehat{D}^*_{j|\mathcal{C}}\right)_{j\in\mathcal{C}^c} \xrightarrow{D} \left(\sum_{i=1}^{(H-1)(d+1)} z^2_{ij} + \sum_{i=1}^{H-1} \widetilde{z}^2_{ij}\right)_{j\in\mathcal{C}^c}, \tag{2.10}$$

where $\mathbf{z}_i$'s and $\widetilde{\mathbf{z}}_i$'s are mutually independent with

$$\mathbf{z}_i = (z_{ij})_{j\in\mathcal{C}^c} \sim \text{MVN}\left(\mathbf{0}, \left[\text{Corr}\left(X_j, X_k|\mathbf{X}_\mathcal{C}\right)\right]_{j,k\in\mathcal{C}^c}\right),$$

and

$$\widetilde{\mathbf{z}}_i = (\widetilde{z}_{ij})_{j\in\mathcal{C}^c} \sim \text{MVN}\left(\mathbf{0}, \left[\text{Corr}^2\left(X_j, X_k|\mathbf{X}_\mathcal{C}\right)\right]_{j,k\in\mathcal{C}^c}\right).$$

We have, as $n \to \infty$,

$$\begin{aligned} & \widehat{D}^*_{j|\mathcal{C}} \xrightarrow{a.s.} D^*_{j|\mathcal{C}} \\ = \; & \log\left(1 + \frac{\text{Var}(M_j) - \text{Cov}\left(M_j, \mathbf{X}_\mathcal{C}\right)\left[\text{Cov}\left(\mathbf{X}_\mathcal{C}\right)\right]^{-1}\text{Cov}\left(M_j, \mathbf{X}_\mathcal{C}\right)^T}{\mathbb{E}\left(V_j\right)}\right) \\ & + \log\mathbb{E}\left(V_j\right) - \mathbb{E}\log\left(V_j\right), \end{aligned}$$

where $M_j = \mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}, S(Y)\right)$, $V_j = \text{Var}\left(X_j|\mathbf{X}_\mathcal{C}, S(Y)\right)$ and $S(Y) = h$ when $Y \in S_h$ $(1 \leq h \leq H)$. According to the Cauchy-Schwarz inequality and Jensen's inequality,

$$D^*_{j|\mathcal{C}} = 0 \text{ iff } \mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}, Y \in S_h\right) = \mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}\right), \text{ and } \text{Var}\left(X_j|\mathbf{X}_\mathcal{C}, Y \in S_h\right) = \text{Var}\left(X_j|\mathbf{X}_\mathcal{C}\right),$$

for $1 \leq h \leq H$. That is, the augmented test statistic $\widehat{D}^*_{j|\mathcal{C}}$ almost surely converges to zero if both the conditional mean and the conditional variance of $X_j$ is independent of slice membership $S(Y)$. Proofs on properties of the augmented likelihood-ratio test statistic are delegated to Appendix A.4.

A forward-addition backward-deletion algorithm similar to the stepwise procedure proposed in Section 2.1.1 can be used with the augmented likelihood-ratio test statistic $\widehat{D}^*_{j|\mathcal{C}}$. To investigate the power of the augmented likelihood-ratio test, we introduce the following

concepts.

**Definition 2** (Conditionally Detectable)**.** *We say a collection of predictors indexed by $\mathcal{C}_2$ is conditionally detectable given predictors indexed by $\mathcal{C}_1$ if $\mathcal{C}_2 \cap \mathcal{C}_1 = \emptyset$, and for any set $\mathcal{C}$ satisfying $\mathcal{C}_1 \subset \mathcal{C}$ and $\mathcal{C}_2 \not\subset \mathcal{C}$, there exist constants $\kappa \geq 0$, $\xi_1, \xi_2 > 0$ such that either*

$$\max_{j \in \mathcal{C}^c \cap \mathcal{C}_1} \left[ \frac{Var(M_j) - Cov(M_j, \mathbf{X}_{\mathcal{C}}) \left[Cov(\mathbf{X}_{\mathcal{C}})\right]^{-1} Cov(M_j, \mathbf{X}_{\mathcal{C}})^T}{\mathbb{E}(V_j)} \right] \geq \xi_1 n^{-\kappa}, \tag{2.11}$$

*or*

$$\max_{j \in \mathcal{C}^c \cap \mathcal{C}_1} \left[ \log(\mathbb{E}V_j) - \mathbb{E}\log(V_j) \right] \geq \xi_2 n^{-\kappa}$$

*where $M_j = \mathbb{E}(X_j | \mathbf{X}_{\mathcal{C}}, S(Y))$, $V_j = Var(X_j | \mathbf{X}_{\mathcal{C}}, S(Y))$.*

In other words, if the current selection $\mathcal{C}$ contains $\mathcal{C}_1$, then there always exist detectable predictors conditioning on currently selected variables until we include all the predictors indexed by $\mathcal{C}_2$. A relevant predictor $X_j$ indexed by $j \notin \mathcal{C}_2$ is *not* conditionally detectable given $\mathcal{C}_1$ either because it is highly correlated with some other predictors, or its effect can only be detected when conditioning on predictors that have not been included in $\mathcal{C}_1$. Based on Definition 2, we define *stepwise detectable* recursively as following.

**Definition 3** (Stepwise Detectable)**.** *A collection of predictors indexed by $\mathcal{T}_0$ is said to be 0-level detectable if $\mathbf{X}_{\mathcal{T}_0}$ is conditionally detectable given an empty set, and a collection of predictors indexed by $\mathcal{T}_m$ is said to be $m$-level detectable ($m \geq 1$) if $\mathbf{X}_{\mathcal{T}_m}$ is conditionally detectable given predictors indexed by $\cup_{i=1}^{m-1} \mathcal{T}_i$. Finally, a predictor indexed by $j$ is said to be stepwise detectable if $j \in \cup_{i=1}^{\infty} \mathcal{T}_i$.*

According to Lemma 1 in Appendix A.2, given the same constant $\kappa$, there exists $\xi_1$ such that the set of marginally detectable predictors defined in Definition 1 is contained in $\mathcal{T}_0$, the

set of 0-level detectable predictors. As a result, the definition of stepwise detectable expand the concept of marginally detectable. In Appendix A.5, we will show that by appropriately choosing thresholds, the stepwise procedure will keep adding predictors until all the stepwise detectable predictors have been included.

**Theorem 2.** *Under Condition 1 and Condition 2, if all the relevant predictors indexed by $\mathcal{A}$ in model (2.3) are stepwise detectable with constant $\kappa$, then there exists constant $c^* > 0$ such that as $n \to \infty$,*

$$\Pr\left(\min_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{A}\neq\emptyset}\max_{j\in\mathcal{C}^c}\widehat{D}^*_{j|\mathcal{C}} \geq c^* n^{-\kappa}\right) \to 1, \text{ and } \Pr\left(\max_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{A}=\emptyset}\max_{j\in\mathcal{C}^c}\widehat{D}^*_{j|\mathcal{C}} < \frac{c^*}{2} n^{-\kappa}\right) \to 1.$$

Thus, by appropriately choosing the thresholds, the stepwise procedure based on $\widehat{D}^*_{j|\mathcal{C}}$ is consistent in identifying stepwise detectable predictors.

### 2.1.3 A Sure Independence Screening Strategy

When the dimensionality $p$ is extremely large (*e.g.*, exceeding $n^2$), the performance of the stepwise procedure can be compromised. Therefore, we recommend adding an independence screening step to first reduce the dimensionality from ultra-high to moderately high. A natural choice of test statistic for the independence screening procedure is $\widehat{D}^*_j = \widehat{D}^*_{j|\mathcal{C}}$ with $\mathcal{C} = \emptyset$, that is, the augmented likelihood-ratio test statistic used in the first addition step of the stepwise procedure. If we rank predictors according to $\{\widehat{D}^*_j, 1 \leq j \leq p\}$, then a sure independence screening (SIS) procedure that takes the first $n-1$ or $n/\log(n)$ predictors has a high probability (almost surely) of including the *independently detectable* predictors defined below.

**Definition 4** (Independently Detectable)**.** *We say a predictor $X_j$ is independently detectable if there exist constants $\kappa \geq 0$ and $\xi_1, \xi_2$ such that either*

$$\frac{Var\left(\mathbb{E}\left(X_j|S(Y)\right)\right)}{\mathbb{E}\left(Var\left(X_j|S(Y)\right)\right)} \geq \xi_1 n^{-\kappa}, \tag{2.12}$$

*or*

$$\log \mathbb{E}\left(Var\left(X_j|S(Y)\right)\right) - \mathbb{E}\log\left[Var\left(X_j|S(Y)\right)\right] \geq \xi_2 n^{-\kappa}.$$

Simply put, independently detectable predictors have either different means or different variances across slices. Therefore, in the example (1.2), both $X_1$ and $X_2$ are independently detectable because $\mathrm{Var}(X_1|Y \in S_h)$ and $\mathrm{Var}(X_2|Y \in S_h)$ $(1 \leq h \leq H)$ are different across slices.

In Theorem 3, we proved that the SIS procedure based on $\{\widehat{D}_j^*, 1 \leq j \leq p\}$ almost surely includes the independently detectable predictors under the following condition with ultra-high dimensionality of predictors.

**Condition 3.** $\log(p) = O(n^\gamma)$ *as* $n \to \infty$ *with* $0 < \gamma + 2\kappa < 1$*, where* $\kappa$ *is the same constant as in (2.12). Furthermore, the number of the relevant predictors* $|\mathcal{A}| \leq \xi_0 n^\eta$ *with* $\eta + \kappa < 1$ *and constant* $\xi_0 > 0$*.*

We prove the following theorem in Appendix A.6.

**Theorem 3.** *Under Condition 1 and Condition 3, if all the relevant predictors indexed by $\mathcal{A}$ are independently detectable, then there exist $c > 0$ and $C > 0$ such that*

$$\Pr\left(\min_{j \in \mathcal{A}} \widehat{D}_j^* \geq c n^{-\kappa}\right) \to 1,$$

*and*

$$\Pr\left(\left|\{j : \widehat{D}_j^* \geq cn^{-\kappa}, 1 \leq j \leq p\}\right| \leq Cn^{\kappa+\eta}\right) \to 1.$$

According to Theorem 3, we can first use the SIS procedure to reduce the dimensionality from $p$ to a scale below sample size, say $n/\log(n)$, and then apply the stepwise procedure proposed in the previous sections. Note that predictors that are marginally or stepwise detectable according to Definition 1 and Definition 3 are not necessarily independently detectable. Fan and Lv (2008) advocated an iterative procedure that alternates between a large-scale screening and a moderate-scale variable selection to enhance the performance, which will be discussed in the next section.

## 2.2 Cross-Stitching and Cross-Validation

The simple model (2.3) and the augmented model (2.8) compensate each other in terms of the bias-variance trade-off. Given finite observations, model (2.3) is simpler and more powerful when the response is driven by some linear combinations of predictors, while model (2.8) is useful in detecting more complex relationships such as heteroscedastic or interactive effects. Similarly, the SIS procedure introduced in the previous section can be used with a large number of predictors, but cannot pick up stepwise detectable predictors that have the same marginal distributions across slices. To find a balance between simplicity and detectability, we propose the following cross-stitching strategy:

- Step 0: start with the SIS procedure with currently selected predictors $\mathcal{C} = \emptyset$;
- Step 1: select predictors by using the stepwise procedure with addition and deletion steps based on $\widehat{D}_{j|\mathcal{C}}$ in (2.4) and add the selected predictors into $\mathcal{C}$;

- Step 2: select predictors by using the stepwise procedure with addition and deletion steps based on $\widehat{D}^*_{j|\mathcal{C}}$ in (2.9) and add the selected predictors into $\mathcal{C}$;

- Step 3: run the SIS procedure on the remaining predictors conditioning on the current selection $\mathcal{C}$, and iterate Step 1-3 until no more predictors are selected.

We name the proposed procedure *Sliced Inverse Regression for Interaction Detection*, or SIRI for short. A flowchart of the SIRI procedure is illustrated in Figure 2.1.

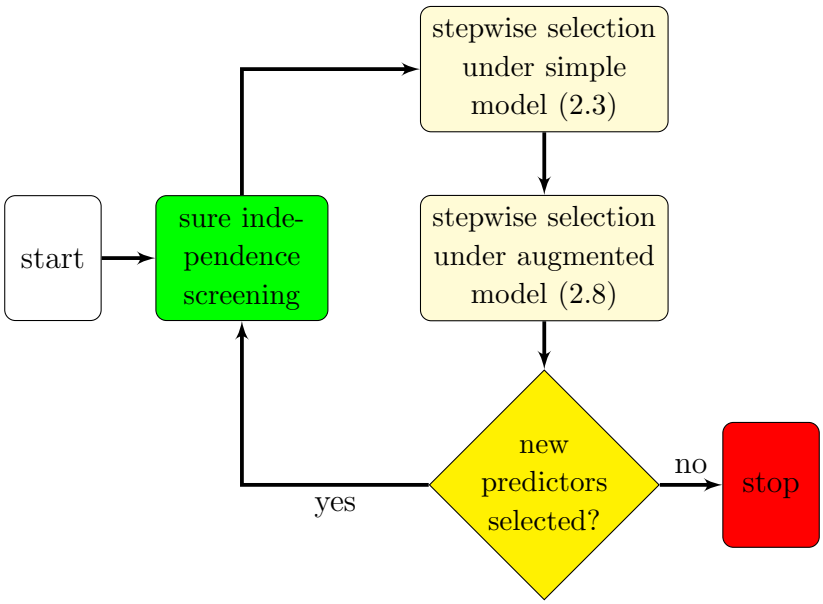


Figure 2.1: Flowchart of SIRI

In the addition step of the stepwise procedure, instead of selecting the variable from $j \in \mathcal{C}^c$ with the maximum value of $\widehat{D}_{j|\mathcal{C}}$ (or $\widehat{D}^*_{j|\mathcal{C}}$), we may also sequentially add variables with $\widehat{D}_{j|\mathcal{C}} > \nu_a$ (or $\widehat{D}^*_{j|\mathcal{C}} > \nu^*_a$). Specifically, given thresholds $\nu_a > \nu_d$ and the current set of selected predictors indexed by $\mathcal{C}$, we can modify each iteration of the original stepwise procedure as follows:

- Modified addition step: for each variable $j \in \{1, \ldots, p\}$, let $\mathcal{C} = \mathcal{C} + \{j\}$ if $j \notin \mathcal{C}$ and $\widehat{D}_{j|\mathcal{C}} > \nu_a$.

- Deletion step: find $j_d$ such that $\widehat{D}_{j_d|\mathcal{C}-\{j_d\}} = \min_{j\in\mathcal{C}} \widehat{D}_{j|\mathcal{C}-\{j\}}$; let $\mathcal{C} = \mathcal{C} - \{j_d\}$ if $\widehat{D}_{j_d|\mathcal{C}-\{j_d\}} < \nu_d$.

The stepwise procedure with the modified addition step may use fewer iterations to find all the relevant predictors and will not stop until all the relevant predictors have been included if we choose $\nu_a = cn^{-\kappa}$ in Theorem 1. However, in practice, the performance of the modified procedure depends on the ordering of the variables and is less stable than the original procedure. Since we are less concerned about the computational cost of SIRI, we implement the original addition step in the following study.

There are some implementation issues that we have not discussed so far. First, we need to choose a slicing scheme. If we assume there is a true slicing scheme from which data are generated, we showed in Appendix A.7 that the power of the stepwise procedure tends to increase with a larger number of slices, but there is no gain by further increasing the number of slices once the slicing is already more refined than the true slicing scheme. In practice, the true slicing scheme is usually unknown (except maybe in the case when the response is discrete). When a slicing scheme uses a larger number of slices, the number of observations in each slice decreases, which makes the estimation of parameters in the model less accurate and less stable. We observed from intensive simulation studies that, with a reasonable number of observations in each slice (say 40 or more), a larger number of slices is preferred.

Second, we need to choose the number of effective directions $q$ in model (2.3) and the thresholds in adding and deleting variables. Section 2.1.1 and 2.1.2 characterize the asymptotic distributions and behaviors of stepwise procedures, and provide some theoretical guidelines for choosing the thresholds. However, these theoretical results are not directly usable because: (1) the asymptotic distributions that we derived in (2.5) and (2.10) are for a single

addition or deletion step; (2) the consistency results are valid in asymptotic sense and the rate of increase in dimension relative to sample size is usually unknown. In practice, we propose to use a $K$-fold cross-validation (CV) procedure for selecting thresholds and the number of effective directions $q$.

We will consider two performance measures for $K$-fold cross-validation: classification error (CE) and mean absolute error (AE). Suppose there are $n$ training samples and $m$ testing samples. The $j$th observation ($j = 1, 2, \ldots, m$) in the testing set has response $y_j$ and slice membership $S_{y_j}$ (the slicing scheme is fixed based on training samples). Let $p_j^{(h)} = \Pr_{\widehat{\theta}}(S(y_j) = h|\mathbf{X} = \mathbf{x}_j)$ be the estimated probability that the observation $j$ is from slice $S_h$, where $\widehat{\theta}$ denotes the maximum likelihood estimate of model parameters. The classification error is defined as

$$\text{CE} = \frac{1}{m}\sum_{j=1}^{m} \mathbb{I}\left[S(y_j) \neq \underset{h}{\text{argmax}}\left(p_j^{(h)}\right)\right].$$

We denote the average response of training samples in slice $S_h$ as

$$\bar{y}^{(h)} = \frac{\sum_{i=1}^{n} \mathbb{I}\left[S(y_i) = h\right] y_i}{\sum_{i=1}^{n} \mathbb{I}\left[S(y_i) = h\right]}, h = 1, 2, \ldots, H,$$

The absolute error is defined as

$$\text{AE} = \frac{1}{m}\sum_{j=1}^{m}\left|y_j - \sum_{h=1}^{H} p_j^{(h)}\bar{y}^{(h)}\right|.$$

CE is a more relevant performance measure when the response is categorical or there is a non-smooth functional relationship (e.g., rational functions) between the response and predictors, and AE is a better measure when there is a monotonic and smooth functional relationship between the response and predictors. There are other measures that have compromise fea-

tures between these two measures, such as median absolute deviation, which will not be explored here. We will use CE and AE as performance measures throughout simulation studies, and name the corresponding methods SIRI-AE and SIRI-CE, respectively.

## 2.3 Simulation Studies

In this section, we study the performance of SIRI and other existing methods by simulations. In order to facilitate fair comparisons with other existing methods that are motivated from the forward model perspective, the examples presented here are all generated under forward models, which differs from the inverse model assumptions of SIRI. The setting of the simulation also demonstrates the robustness of SIRI when some of its model assumptions are violated, especially the normal distribution assumption on relevant predictor variables within each slice.

We start with the comparison of independence screening methods in reducing the ultra-high dimensionality while retaining the relevant predictors. Then, we evaluate different variable selection methods under a variety of forward models including linear model, single- and multi-index models and models with different types of interactions.

### 2.3.1 Independence Screening Performance

We first compare the variable screening performance of SIRI with iterative sure independence screening (ISIS) based on correlation learning proposed by Fan and Lv (2008) and sure independence screening based on distance correlation (DC-SIS) proposed by Li et al. (2012). We evaluate the performance using the proportion that relevant predictors are placed among the top $[n/\log(n)]$ predictors ranked by the corresponding method, with larger values

indicating better performance in variable screening.

In the simulation, the predictor variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ were generated from a $p$-variate normal distribution with mean 0 and covariances $\mathrm{Cov}(X_i, X_j) = \rho^{|i-j|}$ for $1 \leq i, j \leq p$. We generate the response variable from the following three scenarios:

$$\begin{aligned}
\text{Scenario 0.1}: &\quad Y = X_2 - \rho X_1 + 0.2X_{100} + \sigma\epsilon, \\
\text{Scenario 0.2}: &\quad Y = X_1 X_2 + \sigma e^{2|X_{100}|}\epsilon, \\
\text{Scenario 0.3}: &\quad Y = \frac{X_{100}}{X_1 + X_2} + \sigma\epsilon,
\end{aligned}$$

where sample size $n = 200$, $\sigma = 0.2$ and $\epsilon$ is $N(0, 1)$ and independent of $\mathbf{X}$. For each scenario, we consider four different settings with dimension $p = 2000$ or 5000 and correlation $\rho = 0.0$ or 0.5. Scenario 0.1 is a linear model with three additive effects. The way $X_1$ is introduced is to make it marginally uncorrelated with the response $Y$ (note that when $\rho = 0.0$, $X_1$ is not a relevant predictor). We added another variable $X_{100}$ that has negligible correlation with $X_1$ and $X_2$ and a very small correlation with the response $Y$. Scenario 0.2 contains an interaction term $X_1X_2$ and a heteroscedatic noise term determined by $X_{100}$. Scenario 0.3 is an example of a rational model with interactions.

Proportions that relevant predictors are predictors are placed among the top $[n/\log(n)]$ by different screening methods are shown in Table 2.1. Under Scenario 0.1 with linear models, we can see that ISIS and DC-SIS had better power than SIRI in detecting variables that are weakly correlated wth the response ($X_{100}$ in this example). When the correlation between predictors $\rho = 0.5$ (Setting 2 and 4), iterative procedures, ISIS and SIRI, were more effective in detecting predictors that are marginally uncorrelated with the response ($X_1$ in this example) compared with DC-SIS. Under Scenario 0.2, ISIS based on linear models

Table 2.1: The proportions that relevant predictors are placed among the top $[n/\log(n)]$ by different screening methods under Scenarios 0.1-0.3 in Section 2.3.1.

| Method | Scenario 0.1 | | | Scenario 0.2 | | | Scenario 0.3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | $X_1$ | $X_2$ | $X_{100}$ | $X_1$ | $X_2$ | $X_{100}$ | $X_1$ | $X_2$ | $X_{100}$ |
| Setting 1: $p = 2000$, $\rho = 0.0$ | | | | | | | | | |
| ISIS | - | 1.00 | 1.00 | 0.02 | 0.01 | 0.46 | 0.00 | 0.00 | 0.09 |
| DC-SIS | - | 1.00 | 0.55 | 0.07 | 0.09 | 1.00 | 0.00 | 0.00 | 0.60 |
| SIRI | - | 1.00 | 0.30 | 0.32 | 0.25 | 0.97 | 1.00 | 0.99 | 1.00 |
| Setting 2: $p = 2000$, $\rho = 0.5$ | | | | | | | | | |
| ISIS | 1.00 | 1.00 | 1.00 | 0.04 | 0.02 | 0.54 | 0.00 | 0.00 | 0.15 |
| DC-SIS | 0.02 | 1.00 | 0.71 | 0.55 | 0.53 | 1.00 | 0.03 | 0.00 | 0.59 |
| SIRI | 1.00 | 1.00 | 0.45 | 0.92 | 0.87 | 0.92 | 1.00 | 1.00 | 1.00 |
| Setting 3: $p = 5000$, $\rho = 0.0$ | | | | | | | | | |
| ISIS | - | 1.00 | 1.00 | 0.02 | 0.00 | 0.43 | 0.00 | 0.00 | 0.06 |
| DC-SIS | - | 1.00 | 0.39 | 0.03 | 0.05 | 1.00 | 0.00 | 0.00 | 0.44 |
| SIRI | - | 1.00 | 0.14 | 0.15 | 0.16 | 0.99 | 0.99 | 1.00 | 1.00 |
| Setting 4: $p = 5000$, $\rho = 0.5$ | | | | | | | | | |
| ISIS | 1.00 | 1.00 | 1.00 | 0.03 | 0.02 | 0.60 | 0.00 | 0.00 | 0.07 |
| DC-SIS | 0.05 | 1.00 | 0.71 | 0.41 | 0.44 | 1.00 | 0.00 | 0.02 | 0.61 |
| SIRI | 1.00 | 1.00 | 0.39 | 0.88 | 0.86 | 0.94 | 0.98 | 1.00 | 0.99 |

failed to detect the interaction term and often misses the predictor in the heteroscedastic noise term. When there are moderate correlations between two predictors $X_1$ and $X_2$ in the interaction term (Setting 2 and 4), DC-SIS picked up $X_1$ and $X_2$ about half of the time. However, when the two predictors are uncorrelated (Setting 1 and 3), DC-SIS failed to detect them most of the time. SIRI outperformed DC-SIS in detecting interactions for both settings with $\rho = 0.0$ and $\rho = 0.5$. Under Scenario 0.3, when there is a rational relationship between the response and the relevant predictors, SIRI significantly outperformed the other two methods in detecting the relevant predictors. Performances of different methods are only slightly affected as we increase the dimension from $p = 2000$ to $p = 5000$.

### 2.3.2 Variable Selection Performance

We further study the variable selection accuracy of SIRI and other existing methods with simulations in identifying relevant predictors and excluding irrelevant predictors. In the following examples, for both SIRI and COP, we implemented a fixed slicing scheme with 5 slices of equal size (*i.e.*, $H = 5$) and used a 10-fold CV procedure to determine the stepwise variable selection thresholds and the number of effective directions $q$ in model (2.3) of Section 2.1.1. Specifically, the number of effective directions $q$ was chosen from $\{0, 1, 2, 3, 4\}$, where $q = 0$ means that we skipped the variable selection step under simple model (2.3) in the iterative procedure described by Figure 2.1. The thresholds in addition and deletion steps were selected from the grid $\{(\nu_{i,a} = \chi^2_{\alpha_i,q}, \nu_{i,d} = \chi^2_{\alpha_i-0.05,q})\}$ for simple model (2.3) and from the grid $\{(\nu^*_{i,a} = \frac{n}{n-H(d+2)}\chi^2_{\alpha_i,(H-1)(d+2)}, \nu^*_{i,d} = \frac{n}{n-H(d+2)}\chi^2_{\alpha_i-0.05,(H-1)(d+2)})\}$ for augmented model (2.8), where $\chi^2_{\alpha,\text{d.f.}}$ is the $100\alpha$th quantile of $\chi^2_{\text{d.f.}}$ and $d = |\mathcal{C}|$ is the number of previously selected predictors. For a given $p$, the dimension of predictors, we chose $\{\alpha_i\} = \{1 - p^{-1}, 1 - 0.5p^{-1}, 1 - 0.1p^{-1}, 1 - 0.05p^{-1}, 1 - 0.01p^{-1}\}$.

The other variable selection methods to be compared with SIRI and COP include Lasso, ISIS-SCAD (SCAD with iterative sure independence screening), and hierNet (Bien et al., 2012), which is a Lasso-like procedure to detect multiplicative interactions between predictors under hierarchical constraints. The R packages glmnet, SIS, COP and hierNet are used to run Lasso, ISIS-SCAD, COP and hierNet, respectively. For Lasso and hierNet, we select the largest regularization parameter with estimated CV error less than or equal to the minimum estimated CV error plus one standard deviation of the estimate. The tuning parameters in SCAD are also selected by CV.

For variable selections under index models, we generated the predictor variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$ from a multivariate normal distribution with mean 0 and covariances

$\text{Cov}(X_i, X_j) = \rho^{|i-j|}$ for $1 \leq i, j \leq p$ , and simulated the response variable according to the following models:

$$\begin{aligned}
&\text{Scenario 1.1}: \quad Y = \boldsymbol{\beta}^T \mathbf{X} + \sigma\epsilon, \quad n = 200, \sigma = 1.0, \rho = 0.5, \\
&\qquad\qquad \boldsymbol{\beta} = (3, 1.5, 2, 2, 2, 2, 2, 2, 0, \ldots, 0), \\
&\text{Scenario 1.2}: \quad Y = \frac{\sum_{j=1}^{3} X_j}{0.5 + (1.5 + \sum_{j=2}^{4} X_j)^2} + \sigma\epsilon, \quad n = 200, \sigma = 0.2, \rho = 0.0, \\
&\text{Scenario 1.3}: \quad Y = \frac{\sigma\epsilon}{1.5 + \sum_{j=1}^{8} X_j}, \quad n = 1000, \sigma = 0.2, \rho = 0.0,
\end{aligned}$$

where $n$ is the number of observations, $p$ is the number of predictors and is set as 1000 here, and the noise $\epsilon$ is independent of $\mathbf{X}$ and follows $N(0, 1)$. Scenario 1.1 is a linear model which involves 8 true predictors and 992 irrelevant predictors. Scenario 1.2, a multi-index model with 4 true predictors, was studied in Li (1991) and Zhong et al. (2012), and there is a non-linear relationship between the response $Y$ and two linear combinations of predictors $X_1 + X_2 + X_3$ and $X_2 + X_3 + X_4$. Scenario 1.3 is a single-index model with 8 true predictors and heteroscedastic noise.

For each simulation setting, we randomly generated 100 data sets each with $n$ observations and applied variable selection methods to each data set. Two quantities, the average number of irrelevant predictors falsely selected as true predictors (which is referred to as FP) and the average number of true predictors falsely excluded as irrelevant predictors (which is referred to as FN), were used to measure the variable selection performance of each method. For example, under Scenario 1.1, the FPs and FNs range from 0 to 992 and from 0 to 8, respectively, with smaller values indicating better accuracies in variable selection. The FP- and FN-values of different methods together with their corresponding standard errors (in brackets) are reported in Table 2.2.

Table 2.2: False positive (FP) and false negative (FN) values of different variable selection methods under Scenario 1.1-1.3.

| Method | Scenario 1.1 | | Scenario 1.2 | | Scenario 1.3 | |
|---|---|---|---|---|---|---|
| | FP $(0, 992)$ | FN $(0, 8)$ | FP $(0, 996)$ | FN $(0, 4)$ | FP $(0, 992)$ | FN $(0, 8)$ |
| Lasso | 0.59 (0.10) | 0.00 (0.00) | 0.08 (0.03) | 1.07 (0.03) | 0.00 (0.00) | 8.00 (0.00) |
| ISIS-SCAD | 0.35 (0.07) | 0.00 (0.00) | 0.60 (0.08) | 1.02 (0.01) | 5.08 (0.65) | 7.97 (0.02) |
| hierNet | 0.59 (0.10) | 0.00 (0.00) | 8.65 (0.36) | 0.93 (0.03) | 7.66 (0.48) | 7.94 (0.02) |
| COP | 0.69 (0.12) | 0.06 (0.03) | 1.84 (0.16) | 0.98 (0.01) | 1.26 (0.13) | 3.32 (0.19) |
| SIRI-AE | 0.01 (0.01) | 0.09 (0.04) | 0.13 (0.04) | 0.07 (0.03) | 0.43 (0.08) | 4.82 (0.27) |
| SIRI-CE | 0.26 (0.05) | 0.08 (0.03) | 0.55 (0.08) | 0.09 (0.03) | 2.02 (0.17) | 0.51 (0.16) |

Under Scenario 1.1, variable selection methods derived from linear models (Lasso, SCAD and hierNet) were able to detect all the relevant predictors (FN=0) with few false positives. On the other hand, COP, SIRI-AE and SIRI-CE missed some (about 10%) relevant predictors while excluded most irrelevant ones (lower FP vaues). The relatively high accuracy of methods developed for linear models is expected under this scenario, because the observations were simulated from a linear relationship. Under Scenario 1.2, Lasso achieved the lowest false positives, but it almost always missed one of the relevant predictor, $X_4$, because of its non-linear relationship with the response. The other methods developed under the linear model assumption suffered from the issue. However, SIRI-AE and SIRI-CE was able to detect most of the four relevant predictors (FN=0.09 and 0.07) with a comparable number of false positives. Under the heteroscedastic model in Scenario 1.3, the methods based on linear models failed to detect relevant predictors most of the time. Among other methods, SIRI-AE achieved the lowest number of false positives (FP=0.43) but missed about half of the relevant predictors (FN=4.82), while SIRI-CE selected most of the relevant predictors (FN=0.51) with a reasonably low false positives (FP=2.02). The performance of COP was in-between SIRI-AE and SIRI-CE with FN=3.32 and FP=1.26. A possible explanation for

the superior performance of SIRI-CE relative to SIRI-AE in this setting is because the generative model under Scenario 1.3 contains a singular point at $\sum_{j=1}^{8} X_j = -1.5$. Since the absolute error is less robust to outliers than the classification error, SIRI-AE is more sensitive to the inclusion of irrelevant predictors and more conservative in selecting predictors.

Next, we consider forward models containing different types of interactions. Predictor variables $X_1, X_2, \ldots, X_p$ were independent and identically distributed $N(0,1)$ random variables, and the response was generated under the following models given the predictors:

$$\begin{aligned}
&\text{Scenario 2.1}: \quad Y = X_1X_2 + \sigma\epsilon, \quad n = 200,\\
&\text{Scenario 2.2}: \quad Y = X_1 + X_1X_2 + X_1X_3 + \sigma\epsilon, \quad n = 200,\\
&\text{Scenario 2.3}: \quad Y = X_1X_2 + X_1X_3 + \sigma\epsilon, \quad n = 200,\\
&\text{Scenario 2.4}: \quad Y = X_1X_2X_3 + \sigma\epsilon, \quad n = 200, 500 \text{ and } 1000,\\
&\text{Scenario 2.5}: \quad Y = X_1^2X_2 + \sigma\epsilon, \quad n = 200,\\
&\text{Scenario 2.6}: \quad Y = \frac{X_1}{X_2 + X_3} + \sigma\epsilon, \quad n = 200,
\end{aligned}$$

where $n$ is the number of observations, $p$ is the number of predictors and is set as 1000 here, $\sigma = 0.2$ and $\epsilon$ is independent of $\mathbf{X}$ and follows $N(0,1)$. Scenario 2.1 and Scenario 2.3 contain predictors with pairwise multiplicative interactions and without main effects. The model under Scenario 2.2 has hierarchical interaction terms ($X_1$ has main effect). The three-way interaction model in Scenario 2.4 was simulated under three settings with different sample sizes: $n = 200$, $n = 500$ and $n = 1000$. Scenario 2.5 contains a quadratic interaction term and Scenario 2.6 has a rational relationship.

Because methods such as Lasso, SCAD and COP are not specifically designed for detecting interactions and are clearly at a disadvantage, we did not directly compare them with

SIRI and hierNet. For the purpose of comparison, we created a benchmark method based on ISIS-SCAD by applying ISIS-SCAD to an expanded set of predictors that includes all the terms up to $k$-way multiplicative interactions. The corresponding method, which we referred to as ISIS-SCAD-$k$, is an *oracle* benchmark under Scenario 2.1-2.4 when responses were generated according to 2-way or 3-way multiplicative interactions. Since DC-SIS as a screening tool has the ability to detect individual predictors under the presence of interactive effects, we also augmented ISIS-SCAD with DC-ISIS and denoted the method as DC-SIS-SCAD-$k$. In DC-SIS-SCAD-$k$, we first used DC-SIS to reduce the number of predictors from $p$ to $[n/\log(n)]$. Then, we expanded the selected predictors to include up to $k$-way multiplicative interactions among them and applied ISIS-SCAD. Because DC-SIS-SCAD-$k$ does not need to consider all the interaction terms among $p$ predictors, it has a huge speed advantage over ISIS-SCAD-$k$ but it may fail to detect all the predictors if the DC-SIS step does not retain all the relevant predictors. The FP- and FN-values (and their standard errors) of different methods including ISIS-SCAD-$k$ and DC-SIS-SCAD-$k$ under various scenarios are shown in Table 2.3, Table 2.4 and Table 2.5, respectively. Note that FP- and FN-values are calculated based on the number of predictors selected by a method, not based on the number of parameters used in building the model. For example, if $X_3$, $X_4$ and $X_3X_4$ all have non-zero coefficients from hierNet under Scenario 2.1, we count the number of false positives as 2, not 3.

Under Scenarios 2.1-2.3 of Table 2.3, the *oracle* benchmark, ISIS-SCAD-2, correctly discovered most of the relevant predictors in the two-way interactions and did not pick up any irrelevant predictor. It is encouraging to see that the performance of the proposed method SIRI-AE was comparable with ISIS-SCAD-2 (in terms of both false positives and false negatives), although SIRI-AE did not assume the knowledge on the generative model. Moreover,

Table 2.3: False positive (FP) and false negative (FN) values of different variable selection methods under Scenario 2.1-2.3.

| Method | Scenario 2.1 | | Scenario 2.2 | | Scenario 2.3 | |
|---|---|---|---|---|---|---|
| | FP $(0, 998)$ | FN $(0, 2)$ | FP $(0, 997)$ | FN $(0, 3)$ | FP $(0, 997)$ | FN $(0, 3)$ |
| ISIS-SCAD-2 | 0.00 (0.00) | 0.00 (0.00) | 0.00 (0.00) | 0.06 (0.04) | 0.00 (0.00) | 0.03 (0.03) |
| DC-SIS-SCAD-2 | 0.00 (0.00) | 0.00 (0.00) | 0.25 (0.09) | 0.11 (0.03) | 1.56 (0.19) | 1.81 (0.11) |
| hierNet | 2.38 (0.33) | 0.00 (0.00) | 6.93 (0.56) | 0.14 (0.05) | 6.98 (0.57) | 0.12 (0.05) |
| SIRI-AE | 0.01 (0.01) | 0.00 (0.00) | 0.02 (0.01) | 0.04 (0.02) | 0.10 (0.04) | 0.11 (0.05) |
| SIRI-CE | 0.76 (0.13) | 0.00 (0.00) | 0.29 (0.06) | 0.10 (0.04) | 0.86 (0.12) | 0.11 (0.05) |

since both ISIS-SCAD-2 and hierNet considered all the pairwise interactions between $p$ predictor variables, they have computational complexity $O(np^2)$ with $p = 1000$ and need much more computational resources compared with SIRI. On average ISIS-SCAD-2 and hierNet are more than 100 times slower than SIRI (see Table 2.6 for running time comparison of different methods). While we can dramatically increase the computational speed by using DC-SIS to screen variables before applying more refined variable selection methods, relevant predictors may be incorrectly filtered by the DC-SIS procedure as shown by DC-SIS-SCAD's higher false negatives under Scenario 2.3 of Table 2.3.

Table 2.4: False positive (FP) and false negative (FN) values of different variable selection methods under Scenario 2.4 with different sample sizes.

| Method | Scenario 2.4 ($n = 200$) | | Scenario 2.4 ($n = 500$) | | Scenario 2.4 ($n = 1000$) | |
|---|---|---|---|---|---|---|
| | FP $(0, 997)$ | FN $(0, 3)$ | FP $(0, 997)$ | FN $(0, 3)$ | FP $(0, 997)$ | FN $(0, 3)$ |
| DC-SIS-SCAD-3 | 0.45 (0.12) | 0.85 (0.12) | 0.00 (0.00) | 0.00 (0.00) | 0.00 (0.00) | 0.00 (0.00) |
| hierNet | 7.22 (0.64) | 2.41 (0.08) | 7.73 (1.17) | 2.38 (0.08) | 4.25 (1.17) | 2.62 (0.06) |
| SIRI-AE | 0.98 (0.12) | 2.27 (0.06) | 0.36 (0.09) | 0.70 (0.07) | 0.21 (0.06) | 0.00 (0.00) |
| SIRI-CE | 1.98 (0.16) | 2.27 (0.07) | 1.96 (0.17) | 0.46 (0.05) | 2.03 (0.19) | 0.00 (0.00) |

Under Scenario 2.4 with three-way interactions, the computational cost prevents us from directly applying ISIS-SCAD-3 to consider all the three-way interaction terms. So we compared the performance of ISIS-SCAD-3 after variable screening using DC-SIS, that is, DC-

SIS-SCAD-3 in Table 2.4. DC-SIS-SCAD-3 performed best under different sample sizes as it assumed the form of the underlying generative model. Among other methods, the performance of SIRI-AE improved relatively to DC-SIS-hierNet as sample size increased. When sample size $n = 1000$, SIRI-AE was able to select all the relevant predictors with very low false positives.

Table 2.5: False positive (FP) and false negative (FN) values of different variable selection methods Scenario 2.5 and 2.6.

| Method | Scenario 2.5 | | Scenario 2.6 | |
|---|---|---|---|---|
| | FP $(0, 998)$ | FN $(0, 2)$ | FP $(0, 997)$ | FN $(0, 3)$ |
| ISIS-SCAD-2 | 0.04 (0.02) | 1.09 (0.04) | 0.00 (0.00) | 3.00 (0.00) |
| DC-SIS-SCAD-2 | 2.38 (0.18) | 0.51 (0.05) | 0.81 (0.16) | 2.96 (0.02) |
| hierNet | 2.42 (0.44) | 0.88 (0.05) | 5.71 (0.59) | 2.91 (0.03) |
| SIRI-AE | 0.08 (0.03) | 0.00 (0.00) | 0.51 (0.11) | 0.00 (0.00) |
| SIRI-CE | 0.88 (0.11) | 0.01 (0.01) | 0.56 (0.11) | 0.00 (0.00) |

Table 2.6: Average running time (in seconds) of different variable selection methods under Scenarios 2.1-2.3, 2.5 and 2.6

| Method | Scenario 2.1 | Scenario 2.2 | Scenario 2.3 | Scenario 2.5 | Scenario 2.6 |
|---|---|---|---|---|---|
| ISIS-SCAD-2 | 13890.86 | 9406.27 | 11581.55 | 10232.31 | 4220.24 |
| DC-SIS-SCAD-2 | 29.24 | 25.77 | 31.90 | 37.03 | 25.68 |
| hierNet | 15213.01 | 26171.28 | 34733.13 | 37312.59 | 27255.16 |
| SIRI | 27.96 | 44.85 | 20.01 | 44.36 | 35.26 |

Simulations in Scenarios 2.1-2.4 were generated under the same model assumption as ISIS-SCAD-$k$ and DC-SIS-SCAD-$k$, which gives them advantage in the comparisons. Under Scenarios 2.5 and 2.6 of Table 2.5, when the generative model goes beyond multiplicative interactions, we can see that SIRI-AE and SIRI-CE significantly outperformed other methods in detecting relevant predictors with low false positives.

## 2.4 Real Data Examples

We applied SIRI to two real data examples. The first example studies the problem of leukemia subtype classification with ultra-high dimensional features. In the second example, we treat gene expression level in embryonic stem cells as a continuous response variables, and are interested in selecting regulatory factors that interact with DNA and other factors to determine expression patterns of genes.

### 2.4.1 Leukemia Subtypes Classification

For the first example, we applied SIRI-CE to select features for the classification of a leukemia data set from high density Affymetrix oligonucleotide arrays (Golub et al., 1999) that have been previously analyzed by Tibshirani et al. (2002) using a nearest shrunken centroid method and by Fan and Lv (2008) using a SIS-SCAD based linear discrimination method (SIS-SCAD-LD). The data set consists of 7129 genes and 72 samples from two classes: ALL (acute lymphocytic leukemia) with 47 samples and AML (acute mylogenous leukemia) with 25 samples. The data set was divided into a training set of 38 samples (27 in class ALL and 11 in class AML) and a test set of 34 samples (20 in class ALL and 14 in class AML).

Table 2.7: Leukemia classification results

| Method | Training error | Test error | Number of genes |
|---|---|---|---|
| SIRI-CE | 0/38 | 1/34 | 8 |
| SIS-SCAD-LD | 0/38 | 1/34 | 16 |
| Nearest Shrunken Centroid | 1/38 | 2/34 | 21 |

The classification results of SIRI-CE, SIS-SCAD-LD and nearest shrunken centroids method are shown in Table 2.7. The results of SIS-SCAD-LD and the nearest shrunken

centroids method were extracted from Fan and Lv (2008) and Tibshirani et al. (2002), respectively. SIRI-CE and SIS-SCAD-LD both made no training error and one test error, whereas the nearest shrunken centroids method made one training error and two test errors. Comparing with SIS-SCAD-LD, SIRI used a smaller number of genes (8 genes) to achieve the same classification accuracy.

### 2.4.2 Identifying Regulating Factors in Embryonic Stem Cells

The mouse embryonic stem cells (ESCs) data set has previously been analyzed by Zhong et al. (2012) to identify important transcription factors (TFs) for regulating the expression of genes. The response variable, expression levels of 12408 genes, was quantified using RNA-seq technology in mouse ESCs (Cloonan et al., 2008). To understand the ESC development, it is important to identify key regulating TFs, whose binding profiles on cis-regulatory regions are associated with corresponding gene expression levels. To extract features that are associated with potential gene regulating TFs, Chen et al. (2008a) performed ChIP-seq experiments on 12 TFs that are known to play different roles in ES-cell biology as components of the important signaling pathways, self-renewal regulators, and key reprogramming factors. For each pair of gene and one of these 12 TFs, a score named transcription factor association strength (TFAS), which was defined by summing the binding peaks weighted by their intensities and the distances to transcription start sites Ouyang et al. (2009), was calculated. In addition, Zhong et al. (2012) supplemented the data set with motif matching scores of 300 putative mouse TFs compiled from the TRANSFAC database. The TF motif matching scores were calculated based on the occurrences of TF binding motifs on gene promoter regions that are defined as 1kb upstream regions of the transcritipion start sites(Zhong et al., 2005). The data consists of a $12408 \times 312$ matrix with $(i, j)$th entry representing the score of the $j$th

TF on the $i$th gene's promoter region.

Table 2.8: The ranks of 12 known ES-cell TFs (among 312 predictors) using SIRI-AE and COP

| TF names | Ranks | |
|---|---|---|
| | SIRI-AE | COP |
| E2F1 | 1 | 1 |
| ZFX | 3 | 3 |
| MYCN | 4 | 10 |
| KLF4 | 5 | 19 |
| MYC | 6 | - |
| ESRRB | 8 | - |
| OCT4 | 9 | 11 |
| TCFCP2L1 | 10 | 36 |
| NANOG | 14 | - |
| STAT3 | 17 | 20 |
| SOX2 | 18 | - |
| SMAD1 | 32 | 13 |

In Zhong et al. (2012), the method COP selected a total of 42 predictors, which include 8 out of 12 TFASs and 34 out of 300 TF motif scores. Here, we used SIRI-AE to re-analyze the mouse ESCs data set and selected 34 predictors, which include all the 12 TFASs and 22 TF motif matching scores. The relative ranks of 12 TFASs from SIRI-AE and COP are shown in Table 2.8. Among the top-10 TFs ranked by SIRI-AE, 8 of them are known ES-cell TFs. SIRI-AE is also able to identify NANOG and SOX that are generally believed to be the master ESC regulators but were missed in the results of COP. A further study of the top-ranked TFs whose roles in ES cells have not been studied, such as AP2ALPHA and AP2GAMMA (ranked 11 and 12 by SIRI), PAX4 (ranked 19 by SIRI) and SP1 (ranked 22 by SIRI), could help us better understand transcriptional regulatory networks in embryonic stem cells.

## 2.5 Discussion

We studied the problem of variable selection in high dimensions from an inverse modeling perspective. The contributions of the proposed procedure that we named SIRI is twofold. First, it is effective and computationally efficient in detecting interactions. Combined with independence screening, SIRI can be used to attack the problem of ultra-high dimensionality when the number of predictors in the model is much larger than the number of observations. Second, SIRI does not impose any specific assumption on the relationship between the response and predictors, and is a powerful tool for variable selections beyond linear models and detecting predictors with unknown form of interaction effects. As a trade-off, SIRI imposes a few assumptions on the distribution of predictors. As demonstrated in our simulation studies, SIRI has competitive performance when the generative model is different from the inverse model assumption. However, we found that SIRI is not very robust against extreme outliers on values of predictors. Data preprocessing, such as quantile normalization, may be helpful when extreme outliers are presented from exploratory analysis.

Like other stepwise procedures such as linear stepwise regression, SIRI may encounter issues that are typical to stepwise variable selection methods as discussed in Miller (1984). Imagine a simple classification example with two relevant predictors having the same mean and variance, but different correlations in two classes. Then, the predictors are undetectable to any stepwise procedure that selects one variable at a time. In less extreme scenarios, when relevant predictors have weak marginal effects but strong joint effects, iterative sampling procedures such as Gibbs sampling could be more powerful than stepwise procedures like SIRI. This motivates us to further study the problem of variable selection from a Bayesian perspective.

Finally, inverse models are not substitutes but complements to forward models. When a specific form is derived from solid scientific arguments, a forward perspective that treats the distribution of predictors as a nuisance can be more powerful in building predictive models. Depending on one's research questions and objectives, it may be helpful to alternate between the two perspectives in analyzing and interpreting data.

# Chapter 3

# Detecting Pleiotropic and Epistatic Effects in eQTL Study

Expression quantitative trait loci (eQTLs) are genomic loci that regulate expression levels of genes. By assaying gene expression and genetic variation, *e.g.*, single nucleotide polymorphisms (SNPs) or copy number variations (CNVs), simultaneously in segregating populations, scientists wish to correlate variation in the gene expression with genomic sequence variation. In such cases we say that the expression of the query gene(s) is linked to or maps to the corresponding DNA region. One justification for studying genetics of gene expression is that transcript abundance may act as an intermediate phenotype between genomic sequence variation and more complex whole-body phenotypes. Results from eQTL studies have been used for identifying hot spots (Brem et al., 2002; Schadt et al., 2003; Morley et al., 2004; Bystrykh et al., 2005; Chesler et al., 2005; Hubner et al., 2005; Lan et al., 2006), constructing causal networks (Zhu et al., 2004; Bing and Hoeschele, 2005; Chesler et al., 2005; Li et al., 2005; Schadt et al., 2005; Zhu et al., 2008), prioritizing lists of candidate genes for

clinical traits (Bystrykh et al., 2005; Hubner et al., 2005; Schadt et al., 2005) and elucidating subclasses of clinical phenotypes (Schadt et al., 2003; Bystrykh et al., 2005).

## 3.1 Traditional eQTL Mapping and Bayesian Partition Method

Traditional eQTL studies are based on linear regression models (Lander and Botstein, 1989) in which each trait variable is regressed against each marker variable. The $p$-value of the regression slope is reported as a measure of significance for the association. In the context of multiple traits and markers, procedures such as false discovery rate (FDR) controls (Benjamini and Hochberg, 1995; Storey and Tibshirani, 2003) can be used to control family-wise error rates. Despite the success of regression approaches in detecting a single eQTL, a number of challenging problems remain. First, these methods can not easily discover *epistatic effects*, the joint effect of multiple markers. Storey et al. (2005) developed a step-wise regression method to search for pairs of markers. This procedure, however, tends to miss eQTL pairs with small marginal effects but a strong interaction effect. Second, there are often strong correlations among expression levels for groups of genes (called *gene modules*), partially reflecting co-regulation of genes in biological pathways that may respond to common genetic loci and environmental perturbations (Schadt et al., 2003; Yvert et al., 2003; Chen et al., 2008b; Schadt et al., 2008; Zhu et al., 2008). Previous findings of eQTL *hot spots*, *i.e.*, loci affecting a larger number of expression traits than expected by chance, and their biological implications further enhance this notion and highlight the biological importance of finding such *pleiotropic effects*. Mapping genetic loci for multiple traits simultaneously has also been shown to be more powerful than mapping single traits at a time (Jiang and Zeng,

1995). Although for a known small set of correlated traits, one can conduct QTL mapping for the principal components (Mangin et al., 1998), this type of method becomes ineffective when the set size is moderately large or one has to enumerate all possible subsets. An alternative approach is to identify subsets of genes by a clustering method in the first stage, and then fit mixture models to clusters of genes (Kendziorski et al., 2006) or linear regression by treating genes as multivariate responses (Chun and Keleş, 2009). The eQTL mapping then depends on whether the clustering method can find the right number of clusters and the right gene partitions.

The problem of searching for eQTLs can be viewed as a variable selection problem, selecting on both the predictors (genotypes of genetic markers) and the responses (gene expression), including also multi-way interactions among the predictors. Variable selection in regression modeling is a long-standing problem in statistics, especially in analyzing high-dimensional and high-throughput data. Traditional variable selection methods, from which most of the aforementioned methods are derived, focus on the forward modeling perspective, *i.e.*, predictive modeling for the conditional distribution of response(s) $Y$ given predictors $\mathbf{X}$. Our goal here is to detect nontrivial joint effects of subsets of predictors on the response vector. The traditional approaches are therefore rather cumbersome to use since it needs to (a) specify how multiple predictors interact (*e.g.*, a multiplicative effect), and (b) include all possible interaction terms as candidates. Taking a different modeling angle from the regression formulation, Zhang and Liu (2007) introduced the Bayesian epistasis association mapping (BEAM) model to detect epistatic interactions in genome-wide case-control studies, where response $Y$ is a binary variable indicating disease status. The BEAM model can be viewed as a generalization of the naïve Bayes (NB) model, which models $\Pr(\mathbf{X}|Y)$ instead of $\Pr(Y|\mathbf{X})$. Motivated by a naïve Bayes modeling perspective, Zhang et al. (2010) developed

a Bayesian partition (BP) model for eQTL studies based on a joint model of gene expression and genetic markers. More specifically, correlated expression traits $\mathbf{Y}$ and their associated set of markers $\mathbf{X}$ are treated as a *module* in the BP model and a latent *individual type* variable $T$ is introduced to decouple $\mathbf{X}$ and $\mathbf{Y}$ by modeling $\Pr(\mathbf{X}|T)$ and $\Pr(\mathbf{Y}|T)$ separately. A Markov Chain Monte Carlo (MCMC) algorithm (Metropolis et al., 1953) was used to search for the module genes and their linked markers. The BP model has been shown to be consistently more powerful than other methods, such as the two-step regression method (Storey et al., 2005), for detecting epistatic interactions and pleiotropic effects in both simulation and large genetics-genomics studies.

However, the original BP model has several limitations on its flexibility and and scalability due to restrictive model assumptions and computational costs. First, the original BP model only allows positively correlated genes to be selected into the same module and cannot capture complex gene expression patterns in a module. Second, the joint distribution of all the associated markers in a module is described by a saturated model with exponentially growing complexity, which decreases the power of the original BP model in detecting multi-SNP associations especially for markers that are only marginally associated with a module. Moreover, to account for linkage disequilibrium (LD) among adjacent markers, the original BP model imposed a mutually exclusive condition on marker pairs with correlations exceeding a certain threshold, without explicitly modeling the block-wise dependence structure of markers along the genome. Third, the original BP model suffers from the slow convergence of the MCMC algorithm due to a large number of intermediate parameters that are difficult to be analytically marginalized over (summed or integrated out). Although a parallel tempering scheme had been used to help with the mixing of the MCMC chains, it still requires intensive computational resources. Last but not least, the original BP model

assigns a uniform prior on latent individual type $T$. Although uniform clustering priors or Dirichlet process priors are flexible and common choices in Bayesian unsupervised clustering, they may not be suitable for regression-type problems. Our objective is to extract information from genetic markers $\mathbf{X}$ that can be used to explain variation in the gene expression $\mathbf{Y}$. By modeling $\Pr(\mathbf{X}|T)$ and $\Pr(\mathbf{Y}|T)$ separately and assigning a uniform clustering prior on individual type $T$, the original BP model does not take the asymmetric roles played by $\mathbf{X}$ and $\mathbf{Y}$ into account

In this chapter, we address these issues and augment the Bayesian partition model. Under the Bayesian framework with latent individual types, we introduce additional latent variables that partition genes into positively correlated gene clusters and assign multiple gene clusters into a module. This allows us to capture complex dependence structure among gene expression such as negative co-expression. We also divide genetic markers in a module into independent marker groups modeled separately by saturated multinomial models, which increases our ability to detect weak marginal effects. We further improve the Bayesian partition model by simultaneously modeling the block structure of linkage disequilibrium (LD) and selecting SNPs within blocks that are associated with gene expression, either individually or interactively with other SNPs. By collapsing (integrating out) intermediate parameters in the hierarchical model and using dynamic programming to marginalize over LD block partitions, we greatly improve the computational efficiency and the convergence of the MCMC chains. Moreover, by introducing a prior that assigns individual type partitions based on a sorted list (e.g., ranks of gene expression), we ensure that the partition of individuals is primarily determined by variation in gene expression. An efficient dynamic programming algorithm is developed to sample individual types under the proposed prior.

This chapter is organized as following: we start Section 3.2 with a simplified Bayesian

partition model to detect epistatic effects with a single quantitative trait, and introduce the sequential partition prior. In Section 3.3, we describe the full Bayesian partition model of gene modules with pleiotropic effects, which extends the simple model and improves the original BP model. We further discuss the initialization and implementation of the MCMC algorithm designed to obtain posterior samples on parameters of interest in Section 3.4. Simulation studies that compare the augmented BP model with the original BP model and regression-based methods are presented in Section 3.5. In Section 3.6, we illustrate our method on a yeast eQTL data set. We conclude this chapter with a few discussions.

## 3.2 Simple Partition Model with QTL

Suppose $Y_i$ is a quantitative trait, such as the standardized expression level of a single gene, $X_{i,k}$ is the genotype of genetic marker (SNP) $k \in \{1, 2, \ldots, p\}$, and $T_i$ is the unobserved individual type for individual $i \in \{1, 2, \ldots, N\}$. The observed values of $Y_i$ and $X_{i,k}$ are denoted as $y_i$ and $x_{i,k}$. A genetic marker indexed by $k$ is a quantitative trait locus (QTL) that is linked to the trait if indicator $I_k = 1$ and otherwise $I_k = 0$. We denote $Y = \{Y_i\}_{1 \le i \le N}$, $\mathbf{X} = \{X_{i,k} : 1 \le i \le N, 1 \le k \le p\}$ and $\mathbf{X}_{\mathcal{M}} = \{X_{i,k} : 1 \le i \le N, k \in \mathcal{M}\}$ given a marker index set $\mathcal{M}$. Let $\mathcal{A} = \{k : I_k = 1\}$ be the index set of relevant markers. The distribution of gene expression and genotypes of relevant genetic markers are independent conditioning on individual types $T = \{T_i\}_{1 \le i \le N}$, denoted as $\Pr(Y|T)$ and $\Pr(\mathbf{X}_{\mathcal{A}}|T)$, respectively, and the conditional distribution of genotypes of irrelevant markers given relevant ones is independent of $T$, denoted as $\Pr_0(\mathbf{X}_{\mathcal{A}^c}|\mathbf{X}_{\mathcal{A}})$. Given independent prior distributions on individual types $T$

and indicators $I = \{I_k\}_{1 \le k \le p}$, $\pi(T)$ and $\pi(I)$, we can write down their posterior distributions:

$$\begin{aligned} & \Pr(T, I|Y, \mathbf{X}) \propto \pi(T)\pi(I)\Pr(Y|T)\Pr(\mathbf{X}_{\mathcal{A}}|T)\Pr_0(\mathbf{X}_{\mathcal{A}^c}|\mathbf{X}_{\mathcal{A}}) \qquad (3.1) \\ \propto \; & \pi(T)\pi(I)\Pr(Y|T)\Pr(\mathbf{X}_{\mathcal{A}}|T)\frac{\Pr_0(\mathbf{X})}{\Pr_0(\mathbf{X}_{\mathcal{A}})} \\ \propto \; & \pi(T)\pi(I)\Pr(Y|T)\frac{\Pr(\mathbf{X}_{\mathcal{A}}|T)}{\Pr_0(\mathbf{X}_{\mathcal{A}})}, \end{aligned}$$

where $\Pr_0(\mathbf{X}_{\mathcal{A}})$ is the background probability of observing the genotypes of markers in $\mathcal{A}$ and is independent of individual type $T$, and $\Pr_0(\mathbf{X})$ is a constant that does not depend on individual types $T$ or indicators of relevant predictors $I$.

Two issues arise if we chose $\pi(T)$ to be a uniform prior or a Dirichlet process prior. First, the MCMC sampling chain often converges very slowly because of the *label switching* issue, *i.e.*, it moves too infrequently between the symmetric modes corresponding to permutations of individual type labels. Second, because the data consists of a single gene and an overwhelming number of genetic markers, if we allow for all possible partitions of individuals *a priori*, the inference of individual types is more likely to be driven by patterns in genotypes of genetic markers, instead of variation in gene expression. Without taking into account the asymmetrical roles of gene expression and genetic markers, we will not achieve our goal of using genetic information to explain variation in gene expression.

To motivate our choice of the prior for $T$, we first start with a simple case without considering genotypes of genetic markers. In particular, we assume a mixture Gaussian model:

$$Y_i|T_i = t \sim N(\mu_t, \sigma^2), \text{ and } \mu_t \sim N(0, \sigma^2/\kappa),$$

given variance parameter $\sigma^2$ and hyper-parameter $\kappa$. Our goal is to infer unobserved indi-

vidual types based on observations $\{y_i\}_{1\le i\le N}$. Given $\sigma^2$, $\kappa$ and a prior probability $\pi_T$ for adding an individual type, the *maximum a posteriori* (MAP) estimate of individual types can be obtained by the following dynamic slicing procedure:

- Rank individuals according to $\{y_i\}_{1\le i\le N}$ (for simplicity, we assume $y_i$'s have already been sorted such that $y_1 < y_2 < \ldots < y_N$).

- Fill in entries of two tables $[M_j]_{0\le j\le N}$ ($M_0 \equiv 1$) and $[N_j]_{1\le j\le N}$ recursively by increasing order $j = 1, 2, \ldots, N$,

$$\begin{aligned} M_j &= \max_{i\in\{0,1,2,\ldots,j-1\}} \left(\frac{\pi_T}{1-\pi_T} M_i U_{i+1,j}\right), \text{ and} \\ N_j &= \operatorname*{argmax}_{i\in\{0,1,2,\ldots,j-1\}} \left(\frac{\pi_T}{1-\pi_T} M_i U_{i+1,j}\right), \end{aligned} \tag{3.2}$$

  where

$$U_{i+1,j} = \sqrt{\frac{\kappa}{j-i+\kappa}} \exp\left(\frac{\kappa\left(\sum_{k=i+1}^{j} y_k\right)^2}{2\left(j-i+\kappa\right)\sigma^2}\right).$$

- Divide the sorted list of individuals into disjointed slices (individual types) by tracing back the table $[N_j]_{1\le j\le N}$ and choosing the corresponding individual as the end of each slice.

The above procedure is a variant of the *Viterbi algorithm* (Viterbi, 1967). By restricting individual type partitions along the sorted list, we can obtain posterior samples of individual types or equivalently, slice boundaries, by substituting "max" with "$\Sigma$" in the recursive formula and using the *forward summation-backward sampling* algorithm. When $\sigma^2$ is unknown, we can iteratively update individual type $T$ conditioning on $\sigma^2$ and vice versa. The above dynamic slicing procedure provides the intuition and motivation for introducing the

following sequential partition prior.

### 3.2.1 Sequential Partition Prior

Let $R = \{R_i, i = 1, 2, \ldots, N\}$ be a list of individuals sorted according to the quantitative trait (for the ease of description, we assume $R_i = i$ below). Given the list of individual ranks $R$ and a prior probability of $\pi_T$ for adding a new individual type, sequential partition prior assigns probability $(\pi_T)^{N_T-1}(1-\pi_T)^{N-N_T}$ to an individual partition with sequential blocks on the sorted list $R$ $(1 \leq N_T \leq N)$ , and zero probability otherwise. We define $S_i$ to be a slicing indicator, which equals 1 if there is a new partition starting from individual $i$ and 0 otherwise ($S_1 \equiv 1$). Then, the sequential partition prior on $S$ has the following form

$$\pi(S) = \pi(\{S_i\}_{2\leq i\leq N}) = (\pi_T)^{\sum_{i=2}^{N} S_i}(1-\pi_T)^{N-1-\sum_{i=2}^{N} S_i}.$$

Under the sequential partition prior, there is a one to one correspondence between slicing indicators $S = \{S_i\}_{1\leq i\leq N}$ and individual type partition $T = \{T_i\}_{1\leq i\leq N}$.

By simultaneously modeling gene expression and genotypes of relevant genetic markers, we can use the forward-summation backward-sampling algorithm analogous to (3.2) to directly sample from the posterior of $S$ (and hence $T$). To be more specific, we can sum over all possible individual partitions under the sequential partition prior by filling entries of the table $[W_j]_{0\leq j\leq N}$ ($W_0 \equiv 1$) recursively,

$$W_j = \sum_{i=0}^{j-1}\left(\frac{\pi_T}{1-\pi_T} W_i U_{i+1,j} V_{i+1,j}\right), \tag{3.3}$$

where $U_{i+1,j}$ is given in (3.2) and

$$V_{i+1,j} = \Pr_1 \left( \{ x_{s,k} : i+1 \leq s \leq j, k \in \mathcal{A} \} \right)$$

is the probability of observing the genotypes of individuals $\{i+1, i+2, \ldots, j\}$, assuming that individuals $\{i+1, i+2, \ldots, j\}$ are from the same individual type and markers in $\mathcal{A}$ are relevant (the explicit formula is given by (3.5) in the next section).

By marginalizing over individual types $T$ and conditioning on individual ranks $R$ and variance parameter $\sigma^2$, the joint probability of gene expression $Y$ and genotypes of relevant genetic markers $\mathbf{X}_{\mathcal{A}}$ can be obtained from the last entry of the table $[W_j]_{0 \leq j \leq N}$:

$$\Pr\left(Y, \mathbf{X}_{\mathcal{A}} | R, \sigma_c^2\right) = W_N (1-\pi_T)^N \left(2\pi\sigma^2\right)^{\frac{N}{2}} \exp\left( -\frac{\sum_{i=1}^N y_i^2}{2\sigma^2} \right). \tag{3.4}$$

Then, to sample individual type partition $T = \{T_i\}_{1 \leq i \leq N}$, or equivalently $S = \{S_i\}_{1 \leq i \leq N}$, we can use backward sampling based on the table $[W_j]_{0 \leq j \leq N}$.

An appropriate choice of sequential partition prior depends on the sorted list of individuals. In the following sections, we will discuss the issue of choosing individual ranks $R = \{R_i\}_{1 \leq i \leq N}$ when we need to model multiple genes with high correlations.

### 3.2.2 Multinomial Model for Genetic Markers

We use the multinomial model from Zhang and Liu (2007) to model genotypes of genetic markers. In order to make this thesis self-sufficient, we include a detailed description of the model here.

We denote $X_{i,\mathcal{A}} = \{X_{i,k} : k \in \mathcal{A}\}$. Given the individual type $T_i = t$, we assume:

$$X_{i,\mathcal{A}}|T_i = t \sim \text{Multinomial}\left(1, \theta_t\right), \text{ and } \theta_t \sim \text{Dirichlet}\left(\frac{\lambda}{2^{|\mathcal{A}|}}, \ldots, \frac{\lambda}{2^{|\mathcal{A}|}}\right),$$

where $\lambda$ is a hyper-parameter. For the purpose of this study, we assume there are two possible genotypes, *i.e.*, individuals are haplotypes. For diploids, the total number of categories is $3^{|\mathcal{A}|}$ instead of $2^{|\mathcal{A}|}$. After integrating out intermediate parameter $\theta_t$, we can derive the probability of observing $\{x_{i,k} : T_i = t, k \in \mathcal{A}\}$ for individuals from the same individual type $t$:

$$\Pr_1\left(\{x_{i,k} : T_i = t, k \in \mathcal{A}\}\right) = \frac{\Gamma(\lambda)}{\Gamma\left(\lambda + n_t\right)} \prod_{h=1}^{2^{|\mathcal{A}|}} \frac{\Gamma\left(n_t^{(h)} + \frac{\lambda}{2^{|\mathcal{A}|}}\right)}{\Gamma\left(\frac{\lambda}{2^{|\mathcal{A}|}}\right)}, \tag{3.5}$$

where $n_t = |\{T_i = t\}| = \sum_{h=1}^{2^{|\mathcal{A}|}} n_t^{(h)}$ is the number of individuals with type $t$, and $n_t^{(h)}$ is the number of individuals with type $t$ and the $h$th genotype among $2^{|\mathcal{A}|}$ possible genotype combinations from $|\mathcal{A}|$ markers. Then, the probability of observing $\{x_{i,k} : 1 \leq i \leq N, k \in \mathcal{A}\}$ conditioning on individual types $T = \{T_i\}_{1 \leq i \leq N}$ is given by

$$\Pr\left(\mathbf{X}_{\mathcal{A}}|T\right) = \prod_{t=1}^{N_T} \Pr_1\left(\{x_{i,k} : T_i = t, k \in \mathcal{A}\}\right), \tag{3.6}$$

where $N_T$ is the total number of distinct individual types.

### 3.2.3 Block Model of Linkage Disequilibrium

Let $\mathcal{A}$ be an arbitrary set of $s$ markers. We use $l_k$ to denote the genomic coordinate of the $k$th marker in $\mathcal{A}$. Without loss of generality, we assume in this section $\mathcal{A} = \{1, 2, \ldots, s\}$ and $l_k < l_j$ when $k < j$.

To account for linkage disequilibrium (LD) between adjacent markers, we use a block structure model to calculate $\Pr_0(\mathbf{X}_{\mathcal{A}})$, the background probability of observing $\mathbf{X}_{\mathcal{A}}$ assuming markers in $\mathcal{A}$ are not associated with gene expression. We define a block boundary indicator $B_k$ for the $k$th marker, which equals 1 if there is a new LD block starting at a genomic locus between $l_{k-1}$ and $l_k$, and 0 otherwise ($B_1 \equiv 1$). Given $B = \{B_k\}_{1 \le k \le m}$, markers in $\mathcal{A}$ can be partitioned into consecutive blocks. For a block of $(j-k)$ markers $\{k+1, k+2, \ldots, j\}$, we assume

$$O_{k+1,j} = \Pr_0\left(\{x_{i,m} : 1 \le i \le N, k+1 \le m \le j\}\right) = \frac{\Gamma(\lambda)}{\Gamma(\lambda+N)} \prod_{h=1}^{2^{j-k}} \frac{\Gamma\left(n^{(h)} + \frac{\lambda}{2^{j-k}}\right)}{\Gamma\left(\frac{\lambda}{2^{j-k}}\right)},$$

where $\lambda$ is a hyper-parameter, $n^{(h)}$ is the number of individuals with the $h$th genotype among $2^{j-k}$ possible genotype combinations from $(j-i)$ markers and $\sum_{h=1}^{2^{j-k}} n^{(h)} = N$.

Then, we can use the same forward-summation technique similar to (3.3) to sum over all possible block partitions of markers in $\mathcal{A}$. Given $\pi_B$, the prior probability of introducing a new LD block at a genomic locus, we fill in the table $[Q_j]_{0 \le j \le s}$ ($Q_0 \equiv 1$) recursively,

$$Q_j = \sum_{k=0}^{j-1} \frac{1 - (1-\pi_B)^{l_{k+1} - l_k}}{(1-\pi_B)^{l_{k+1} - l_k}} Q_k O_{k+1,j},$$

where $l_0 \equiv l_1$. The background probability of $\mathbf{X}_{\mathcal{A}}$ is given by

$$\Pr_0(\mathbf{X}_{\mathcal{A}}) = (1-\pi_B)^{l_s - l_1} Q_s. \tag{3.7}$$

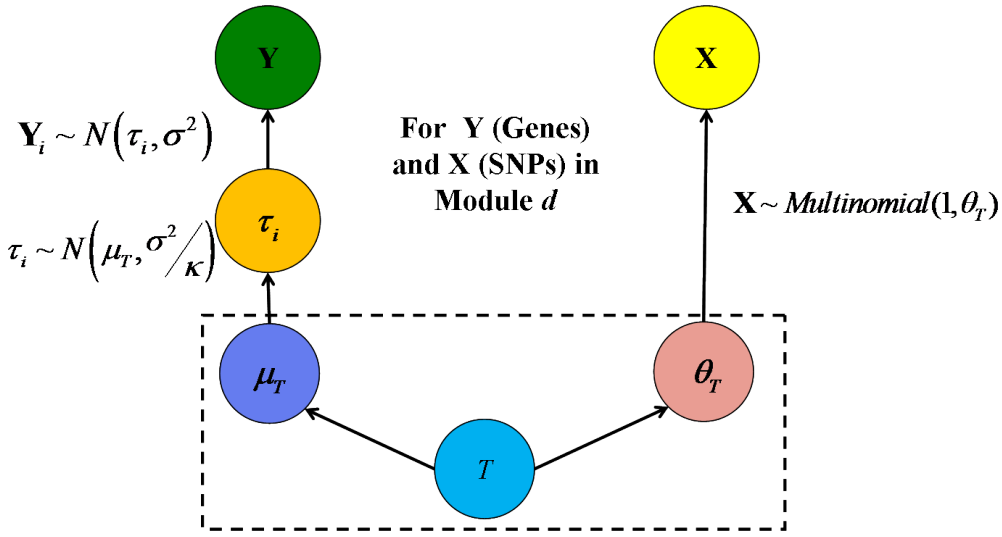


Figure 3.1: An illustration of a module with a single gene cluster denoted as $\mathbf{Y}$ and a single marker group denoted as $\mathbf{X}$. $T$ is the hidden individual type, $\theta_T$ is the parameter for multinomial model of genetic markers described in Section 3.2.2, and $\mu_T$ and $\tau_i$ are parameters for hierarchical model of genes described in Section 3.3.1.

## 3.3 Augmented Partition Model with Gene Modules

We define a module as a set of gene expression traits and a set of genetic markers such that the variation of the gene expression traits is associated with the genotypes of genetic markers. This association between multiple genes and markers is characterized by a latent variable $T$ of individual types. Gene expression traits and genetic markers are conditionally independent given the individual type partition, as illustrated in Figure 3.1. The latent individual type can be viewed as representing a certain combination of markers that induces changes in expression of a certain set of genes. The goal of the Bayesian partition method is to simultaneously partition gene expression traits and genetic markers into modules.

### 3.3.1 Hierarchical Model of Gene Clusters

Here we extend the model of a single gene in Section 3.2 to multiple gene expression traits. Let $Y_{i,j}$ be the quantile normalized and standardized expression level of gene $j \in \{1, 2, \ldots, q\}$ for individual $i \in \{1, 2, \ldots, N\}$ with individual type $T_i = t$. The observed value of $Y_{i,j}$ is denoted as $y_{i,j}$. We further divide genes into clusters such that genes within the same cluster are co-expressed with positive correlations. Give $C_j$, the cluster membership of gene $j$, we assume the following hierarchical model:

$$Y_{i,j}|C_j = c \sim N(\tau_{i,c}, \sigma_c^2), \text{ and } \tau_{i,c}|T_i = t \sim N\left(\mu_{t,c}, \sigma_c^2/\kappa_1\right), \tag{3.8}$$

where $\mu_{t,c} \sim N\left(0, \sigma_c^2/\kappa_2\right)$, $\sigma_c^2 \sim \text{Inv-}\chi^2\left(\nu_0, \sigma_0^2\right)$ and $\kappa_1$, $\kappa_2$, $\nu_0$ and $\sigma_0$ are hyper-parameters. We denote the set of co-expressed genes in cluster $c$ as $\mathcal{K}_c = \{C_j = c\}$ and their expression $\mathbf{Y}_{\mathcal{K}_c} = \{Y_{i,j} : C_j = c, 1 \le i \le N\}$. After integrating out intermediate parameters, we can derive the conditional distribution of $\mathbf{Y}_{\mathcal{K}_c}$ given an individual type partition $T = \{T_i\}_{1 \le i \le N}$,

$$\Pr\left(\mathbf{Y}_{\mathcal{K}_c}|T, \sigma_c^2\right) = \left(2\pi\sigma_c^2\right)^{\frac{Nn_c}{2}} Z_c \exp\left(-\frac{S_c^2}{2\sigma_c^2}\right), \tag{3.9}$$

with

$$Z_c = \left(\sqrt{\frac{\kappa_1}{n_c + \kappa_1}}\right)^N \left(\prod_{t=1}^{N_T} \sqrt{\frac{(n_c + \kappa_1)\kappa_2}{(n_c + \kappa_1)\kappa_2 + n_c n_t \kappa_1}}\right),$$

and

$$S_c^2 = \sum_{i=1}^{N} \left[\sum_{C_j = c} y_{i,j}^2 - \frac{\left(\sum_{C_j=c} y_{i,j}\right)^2}{n_c + \kappa_1}\right] - \sum_{t=1}^{N_T} \frac{\kappa_1^2 \left(\sum_{T_i=t} \sum_{C_j=c} y_{i,j}\right)^2}{(n_c + \kappa_1)\left[(n_c + \kappa_1)\kappa_2 + n_c n_t \kappa_1\right]},$$

where $n_c = |\mathcal{K}_c|$ is the number of genes in cluster $c$, $n_t = |\{T_i = t\}|$ is the number of individuals with type $t$ and $N_T$ is the number of distinct individual types in $T$. We can further integrate out $\sigma_c^2$ given its conjugate prior Inv-$\chi^2(\nu_0, \sigma_0^2)$:

$$\Pr(\mathbf{Y}_{\mathcal{K}_c}|T) = Z_c \frac{\Gamma\left(\frac{Nn_c+\nu_0}{2}\right)(\nu_0\sigma_0^2)^{\frac{\nu_0}{2}}}{[\Gamma(1/2)]^{Nn_c}\,\Gamma(\nu_0/2)\,(S_c^2+\nu_0\sigma_0^2)^{\frac{Nn_c+\nu_0}{2}}}. \tag{3.10}$$

We denote the number of distinct gene clusters as $K$, which is assumed to be fixed. For genes not belonging to any cluster, that is, $\{j : C_j = 0\}$, we assume their standardized expression levels follow independent standard normal distributions, denoted as $\Pr_0\left(\mathbf{Y}_{\{j:C_j=0\}}\right)$.

### 3.3.2 Partition Model with Single Module

To assign gene clusters into modules, for each gene cluster $\mathcal{K}_c$ $(1 \le c \le K)$, we denote its module membership as $J_c$, which equals $d$ if the gene cluster belongs to the eQTL module indexed by $d$ and 0 if the gene cluster does not belong to any module. In addition, we redefine the association indicator for the $k$th genetic marker $(1 \le k \le p)$, and $I_{k,d} = 1$ if the marker is associated with the module indexed by $d$ and $I_{k,d} = 0$ otherwise. An eQTL module indexed by $d$ consists of a family of gene clusters $\{\mathcal{K}_c : J_c = d\}$ and a set of genetic markers $\mathcal{A}_d = \{k : I_{k,d} = 1, 1 \le k \le p\}$. The association between gene clusters and genetic markers is characterized by the common latent individual type partition $T_d$.

The set of markers associated with module $d$, $\mathcal{A}_d$, is further divided into non-overlapping marker groups $\{\mathcal{A}_{d,1}, \ldots, \mathcal{A}_{d,M_d}\}$, where $\cup_{m=1}^{M_d}\mathcal{A}_{d,m} = \mathcal{A}_d$ and the number of distinct marker groups $M_d$ in module $d$ is treated as random. We assume that genotypes of different markers

groups are conditionally independent given individual type partition $T_d$, that is,

$$\Pr\left(\mathbf{X}_{\mathcal{A}_d}|T_d\right) = \prod_{m=1}^{M_d} \Pr\left(\mathbf{X}_{\mathcal{A}_{d,m}}|T_d\right) \tag{3.11}$$

where $\Pr\left(\mathbf{X}_{\mathcal{A}_{d,m}}|T_d\right)$ is given by (3.6) in Section 3.2.2.

For the module indexed by $d$, the sequential partition prior on individual type partition $T_d$ is specified with respect to a sorted list of individuals $R_d = \{R_{d,i}\}_{1\leq i\leq N}$, and denoted as $\pi\left(T_d|R_d\right)$. For a module with one gene cluster, $R_d$ can be obtained by sorting individuals according to the average expression levels of genes in the cluster. The choice of $R_d$ for multiple gene clusters in a module will be discussed in Section 3.4.2. For the current discussion, we assume $R_d$ has been chosen. The joint probability of gene expression $\{\mathbf{Y}_{\mathcal{K}_c} : J_c = d\}$, genotypes $\mathbf{X}_{\mathcal{A}_d}$, and individual type partition $T_d$ can be written as

$$\Pr\left(\mathbf{X}_{\mathcal{A}_d}, \{\mathbf{Y}_{\mathcal{K}_c} : J_c = d\}, T_d|R_d\right) \propto \pi(T_d|R_d)\Pr\left(\mathbf{X}_{\mathcal{A}_d}|T_d\right) \prod_{c:J_c=d} \Pr\left(\mathbf{Y}_{\mathcal{K}_c}|T_d\right), \tag{3.12}$$

where $\Pr\left(\mathbf{X}_{\mathcal{A}_d}|T_d\right)$ is given by (3.11) and $\Pr\left(\mathbf{Y}_{\mathcal{K}_c}|T_d\right)$ is given by (3.10). In practice, to obtain posterior samples of individual type partition $T_d$, we can couple the forward-summation backward-sampling technique introduced in Section 3.2.1 and data augmentation by conditioning on variance parameters $\{\sigma_c^2 : J_c = d\}$. The details of the forward-summation step is given in Appendix B.1.

### 3.3.3 Partition Model with Multiple Modules

The full model includes $D$ modules, with each module's probability distribution given by (3.12), and background models of gene expression and genotypes of genetic markers. We

assume $D$ is fixed here. For gene clusters that are not associated with any module, *i.e.*, $\{\mathcal{K}_c : J_c = 0\}$, since we are not interested in their individual type partitions, we simply assume

$$\Pr\left(\{\mathbf{Y}_{\mathcal{K}_c} : J_c = 0\}\right) = \prod_{c:J_c=0} \Pr\left(\mathbf{Y}_{\mathcal{K}_c}|T_0\right), \tag{3.13}$$

where $\Pr\left(\mathbf{Y}_{\mathcal{K}_c}|T_0\right)$ is given by (3.10) with $T_0 = \{T_{0,i} \equiv 1\}_{1\le i\le N}$, that is, all the individuals are from the same individual type.

Besides individual partitions $\{T_d\}_{1\le d\le D}$, parameters of interest include gene clusters $\{\mathcal{K}_c\}_{1\le c\le K}$ and their module memberships $\{J_c\}_{1\le c\le K}$, as well as genetic markers associated with each module $\{\mathcal{A}_d\}_{1\le d\le D}$ and their partitions into conditionally independent groups $\{\mathcal{A}_{d,m}\}_{1\le m\le M_d, 1\le d\le D}$. Let $\pi_C$, $\pi_J$ and $\pi_I$ be the prior probabilities for adding a gene into a cluster, adding a cluster to a module and adding a genetic marker to a module, respectively. Given observed data and ranks of individuals in each module, $\{R_d\}_{1\le d\le N}$, the posterior probability of parameters of interest under the full model can be formally written as

$$\begin{aligned} & \Pr\left(\{\mathcal{K}_c\}_{1\le c\le K}, \{J_c\}_{1\le c\le K}, \{\mathcal{A}_{d,m}\}_{1\le m\le M_d, 1\le d\le D}, \{T_d\}_{1\le d\le D}|\mathbf{X}, \mathbf{Y}, \{R_d\}_{1\le d\le N}\right) \\ \propto\ & \Pr\left(\{\mathbf{Y}_{\mathcal{K}_c} : J_c = 0\}\right) \prod_{d=1}^{D} \frac{\Pr\left(\mathbf{X}_{\mathcal{A}_d}, \{\mathbf{Y}_{\mathcal{K}_c} : J_c = d\}, T_d|R_d\right)}{\Pr_0(\mathbf{X}_{\mathcal{A}_d})} \times \\ & \left(\frac{\pi_C}{1-\pi_C}\right)^{\sum_{c=1}^{K}|\mathcal{K}_c|} \left(\frac{\pi_J}{1-\pi_J}\right)^{\sum_{c=1}^{K}|\{c:J_c>0\}|} \left(\frac{\pi_I}{1-\pi_I}\right)^{\sum_{d=1}^{D}|\mathcal{A}_d|}, \end{aligned} \tag{3.14}$$

where $\Pr\left(\{\mathbf{Y}_{\mathcal{K}_c} : J_c = 0\}\right)$ is given by (3.13), $\Pr\left(\mathbf{X}_{\mathcal{A}_d}, \{\mathbf{Y}_{\mathcal{K}_c} : J_c = d\}, T_d|R_d\right)$ is given by (3.12) and $\Pr_0(\mathbf{X}_{\mathcal{A}_d})$ is given by (3.7)

Until now, we have an implicit assumption that relevant markers in different modules $\mathcal{A}_d$ $(1 \le d \le D)$ are mutually exclusive, that is, $\sum_{d=1}^{D} I_{k,d} \le 1$ for $1 \le k \le p$. This constraint, which was adopted by Zhang et al. (2010), favors modules with a large number of genes

and genetic markers, but the complexity of the module structure also reduces the detection power given moderate sample size in most eQTL studies. For example, if two genetic marker indexed by $k_1$ and $k_2$ are jointly associated with gene cluster $c_1$, and at the same time, the marker indexed by $k_1$ is marginally associated with gene cluster $c_2$. In order to detect both sets of eQTLs, we have to capture gene clusters $c_1$ and $c_2$ in the same module. If the correlations between two gene clusters are weak, then it is likely that we will miss at least one of the eQTLs. In the following studies, we fix the number of modules $D$ and allow a marker indexed by $k$ to be associated with more than one modules without imposing any constraint on $\{I_{k,d}\}_{1\leq d\leq D}$.

## 3.4 MCMC Algorithm and Implementation

### 3.4.1 Choice of Hyper-Parameters

There are several hyper-parameters that need to be specified, including the number of gene clusters $K$, the number of modules $D$, the prior probabilities $(\pi_T, \pi_B, \pi_C, \pi_J, \pi_I)$, and hyper-parameters $(\kappa_1, \kappa_2, \nu_0, \sigma_0^2)$ in the hierarchical model and $\lambda$ in the multinomial model. In practice, we recommend choosing the number of gene clusters $K$ to be moderately large (say 200 to 400) so that we can capture the correlation structure among genes' expression patterns. Given the number of clusters, we use the initialization strategy introduced in the next section to group gene clusters into super-clusters, and then the number of modules $D$ can be chosen based on the number of super-clusters. Priors $(\pi_I, \pi_B, \pi_T)$ should be chosen based on prior knowledge. In the yeast data set (Brem and Kruglyak, 2005; Zhang et al., 2010), we assume there are 6 SNPs associated with each module *a priori*, and set $\pi_I = 6/3000 = 0.002$. We choose $\pi_B = 0.02$ corresponding to about 5 blocks per chromosome,

and $\pi_T = 0.05$ corresponding to about 20 observations in each individual type. We use $\lambda = 1$, the Jeffreys' prior when there are two possible genotypes. Finally, we found that our results are not sensitive to the choice of other hyper-parameters and set $\pi_C = \pi_G = 0.05$ and $\kappa_1 = \kappa_2 = \nu_0 = \sigma_0^2 = 1$.

### 3.4.2 Initialization of Clusters and Ranks

To specify the sequential partition prior, we need to have a sorted list of individuals $R_d$ for each module. First, we initialize $K$ gene clusters based on a hierarchical model of gene expression without considering genotypes or individual type partitions (gene expression levels are positively correlated in each cluster). Within each initialized cluster, we rank individuals by their average expression levels and denote the ranks of individuals in cluster $c$ as $R_c^*$ $(1 \leq c \leq K)$. We define a *super-cluster* to be a collection of correlated gene clusters, assign $K$ gene clusters into super-clusters using hierarchical clustering based on absolute values of Spearman's correlations between $\{R_c^*\}_{1 \leq c \leq K}$. We link gene clusters in a super-cluster to a module $d$ by letting $J_c = d$. Finally, the ranks of individuals in module $d$, $R_d$, is randomly chosen from $\{R_c^* : J_c = d\}$. In practice, we fix the reservoir of possible ranks $\{R_c : J_c = d\}$ after initialization, and sample $R_d$ from $\{R_c^* : J_c = d\}$ by conditioning on other parameters using a MCMC sampling algorithm.

### 3.4.3 MCMC Algorithm

After initialization, we iteratively update parameters of interest according to their posterior distributions through the following steps:

1. Sample indicators assigning genes to clusters, $\{C_j\}_{1 \leq j \leq q}$. For gene $j = 1, 2, \ldots, q$,

conditioning on $\{C_i : i \neq j\}$ and other parameters, sample $C_j$ according to (3.14).

2. Sample indicators assigning gene clusters to modules, $\{J_c\}_{1\leq c\leq K}$. For gene cluster $c = 1, 2, \ldots, K$, conditioning on $\{J_i : i \neq c\}$ and other parameters, sample $J_c$ according to (3.14).

3. For module $d = 1, 2, \ldots, D$, sample indicators for assigning genetic markers to module $d$, $\{I_{k,d}\}_{1\leq k\leq p}$. Given module $d$ and genetic marker $k = 1, 2, \ldots, p$, conditioning on $\{I_{i,d} : i \neq k\}$ and other parameters, sample $I_{k,d}$ according to (3.14).

4. For gene clusters $c$ with $J_c = d$ $(0 \leq d \leq D)$, sample the variance parameter $\sigma_c^2$ from Inv-$\chi^2\left(Nn_c + \nu_0, \frac{S_c^2+\nu_0\sigma_0^2}{Nn_c+\nu_0}\right)$ conditioning on other parameters, where $S_c^2$ is given by (3.9), $n_c$ is the number of genes in cluster $c$.

5. For module $d = 1, , \ldots, D$, sample individual ranks $R_d$ from $\{R_c^* : J_c = d\}$, *i.e.*, reservoir of possible ranks for module $d$, according to (3.12). Then, use the forward-summation backward-sampling algorithm introduced in Section 3.2.1 to sample individual type partition $T_d$ conditioning on $R_d$ and $\{\sigma_c^2 : J_c = d\}$.

## 3.5 Simulation Studies

In this section, we compare the performance of the augmented Bayesian partition model with the original Bayesian partition model and other eQTL methods. The design of the first simulation study is the same as the design in Zhang et al. (2010), where genes in the same module are positively correlated. To mimic more complex gene expression patterns in real data, in the second simulation study, we modify the original design to allow genes in the same module to be either positively or negatively correlated.

We analyze the simulated data sets using four methods: (1) the original Bayesian partition (BP) model proposed by Zhang et al. (2010), referred to as BPM1; (2) the augmented Bayesian partition model developed in this thesis, referred to as BPM2; (3) a two-stage stepwise regression method applied to individual gene expression proposed by Storey et al. (2005), referred to as SR; (4) a two-stage stepwise regression method applied to the first principle component (PC) of expression levels of known genes in each module, referred to as PCA. The SR method has two stages: in the first stage, it identifies the most significant marker for each gene expression trait based on the one-gene-one-marker regression model. It then proceeds to find the next most significant marker conditional on the previous detected marker for each gene. Permutation tests over all genes are carried out in each stage to control the overall false discovery rate (FDR). The PCA method assumes that the *true* genes in each module are known, and serves as an oracle benchmark for the SR method.

### 3.5.1 Simulation with Positively Correlated Genes

According to the simulation design in Zhang et al. (2010), we simulated 120 individuals with 500 binary markers and 1000 expression traits in the context of inbred cross of haploid strains. Given the haploid nature of the segregants, 500 binary markers are equally spaced on 20 chromosomes, each of length 100cM, using the qtl package in R. There are 8 modules (denoted as A,B,. . .,H), each consisting of 40 genes and 2 associated markers, simulated from different epistasis models based on the linear regression framework. The associated markers in each module are randomly selected and do not overlap. Note that the generative models in our simulation studies are different from the posited Bayesian partition model.

To mimic the inter-correlation of the genes in real gene expression data, we first generated a core gene in each module according to the corresponding models depicted in Table 3.1. In

Table 3.1: Simulation design and genetic variance decomposition

| Module | Model[1] | % of Var.[2] | Locus 1[3] | Locus 2[4] | Epistasis[5] |
|---|---|---|---|---|---|
| A | $Y = \beta \mathbb{I}_{x_1=1 \text{ or } x_2=1} + \epsilon$ | 0.166 | 0.345 | 0.342 | 0.313 |
| B | $Y = \beta \mathbb{I}_{x_1=x_2} + \epsilon$ | 0.166 | 0.058 | 0.054 | 0.888 |
| C | $Y = 2\beta \mathbb{I}_{x_1=1 \text{ or } x_2=1} + \beta x_1 x_2 + \epsilon$ | 0.166 | 0.461 | 0.445 | 0.094 |
| D | $Y = \beta \mathbb{I}_{01} + 2\beta \mathbb{I}_{10} + \epsilon$ | 0.166 | 0.119 | 0.116 | 0.765 |
| E | $Y = \beta x_1 + \beta x_1 x_2 + \epsilon$ | 0.171 | 0.749 | 0.138 | 0.113 |
| F | $Y = 2\beta x_1 + \beta x_2 + \epsilon$ | 0.168 | 0.736 | 0.216 | 0.048 |
| G | $Y = 2\beta x_1 + \beta \mathbb{I}_{x_1=x_2} + \epsilon$ | 0.170 | 0.743 | 0.058 | 0.199 |
| H | $Y = 2\beta \mathbb{I}_{01} + 1.5\beta \mathbb{I}_{10} + 0.5\beta \mathbb{I}_{11} + \epsilon$ | 0.165 | 0.135 | 0.053 | 0.812 |

[1]Regression models that were used to generate the core gene in each module.
[2]Average percentage of variation of genes in the module explained by the true model.
[3]Average percentage of genetic variance explained by the first locus.
[4]Average percentage of genetic variance explained by the second locus.
[5]Average percentage of genetic variance explained by epistasis.

each model, $\epsilon \sim N(0, \sigma_e^2)$ represents the environmental noise. The regression coefficient $\beta$ in each model is determined by the corresponding heritability which is defined as $h^2 = \frac{\sigma_s^2 - \sigma_p^2}{\sigma_p^2}$, where $\sigma_s^2$ and $\sigma_p^2$ are the variance among phenotype values in the segregants and the pooled variance among parental measurements, respectively. Because the variance in the parental measurements reflect only measurement error, we set $\sigma_e^2 = \sigma_p^2 = 1$ in the simulation. For example, in module B, $\sigma_s^2 = \frac{\beta^2}{4} + 1$, $h^2 = 0.6$, thus we are able to solve for the value of $\beta = \sqrt{6}$. After generating the core gene, we simulated the gene expression traits in each module independently from a Gaussian model conditional on the core gene so that they have a given average correlation to the core gene. In this simulation study, we fixed heritability at 0.6 for the core gene and the average inter-correlation for genes in the module with the core gene at 0.5 across all eight modules. Finally, we calculated the percentage of variation explained by the true model averaged over all genes in a module as listed in the third column of Table 3.1. For example, for each gene in module B we calculated the sum of squares of the

gene expression for all 120 samples ($SS_{total}$) and the residual sum of squares ($SS_{res}$) within the two sample groups: those with $x_1 = x_2$ and those with $x_1 \neq x_2$. As a result, the percentage of variation explained by the true model for this gene is $1 - \frac{SS_{res}}{SS_{total}}$.

Table 3.2: Genotype means and frequencies for a two-locus model

| | | Locus 2 | | Mean |
|---|---|---|---|---|
| | | $B$ | $b$ | |
| Locus 1 | $A$ | $\mu_{AB}$ $(p_{AB})$ | $\mu_{Ab}$ $(p_{Ab})$ | $\mu_A$ |
| | $a$ | $\mu_{aB}$ $(p_{aB})$ | $\mu_{ab}$ $(p_{ab})$ | $\mu_a$ |
| Mean | | $\mu_B$ | $\mu_b$ | |

To get a better understanding of the signal strength in each module, we divided the total genetic variance for a two-locus model into three components: the genetic variance at locus 1, the genetic variance at locus 2, and the epistatic (interaction) variance using the classical analysis of variance (Fisher, 1919; Cockerham, 1954; Tiwari and Elston, 1997). In Table 3.2, each row displays the genotype at the first locus, and each column indexes the genotype at the second locus. Each cell contains a genotypic mean and its respective frequency in parentheses. Given the genotypic means and frequencies at both loci, one can calculate the total genetic variance ($\sigma^2$):

$$\sigma^2 = p_{AB}(\mu_{AB} - \mu)^2 + p_{Ab}(\mu_{Ab} - \mu)^2 + p_{aB}(\mu_{aB} - \mu)^2 + p_{ab}(\mu_{ab} - \mu)^2,$$

where $\mu = p_{AB}\mu_{AB} + p_{Ab}\mu_{Ab} + p_{aB}\mu_{aB} + p_{ab}\mu_{ab}$. The amounts of genetic variance at locus 1

and locus 2 are:

$$\begin{aligned}\sigma_1^2 &= p_A(\mu_A-\mu)^2+p_a(\mu_a-\mu)^2,\\ \sigma_2^2 &= p_B(\mu_B-\mu)^2+p_b(\mu_b-\mu)^2,\end{aligned}$$

where $\mu_A=\frac{p_{AB}\mu_{AB}+p_{Ab}\mu_{Ab}}{p_{AB}+p_{Ab}}$, $\mu_a=\frac{p_{aB}\mu_{aB}+p_{ab}\mu_{ab}}{p_{aB}+p_{ab}}$, $\mu_B=\frac{p_{AB}\mu_{AB}+p_{aB}\mu_{aB}}{p_{AB}+p_{aB}}$ and $\mu_b=\frac{p_{Ab}\mu_{Ab}+p_{ab}\mu_{ab}}{p_{Ab}+p_{ab}}$. The epistatic variance is defined as the amount of genetic variance not accounted for by the single-locus components. We performed this decomposition for each gene in a module and calculated the average percentage of the single-locus variances and the epistatic variance for each module. These values are listed in the last three columns of Table 3.1.

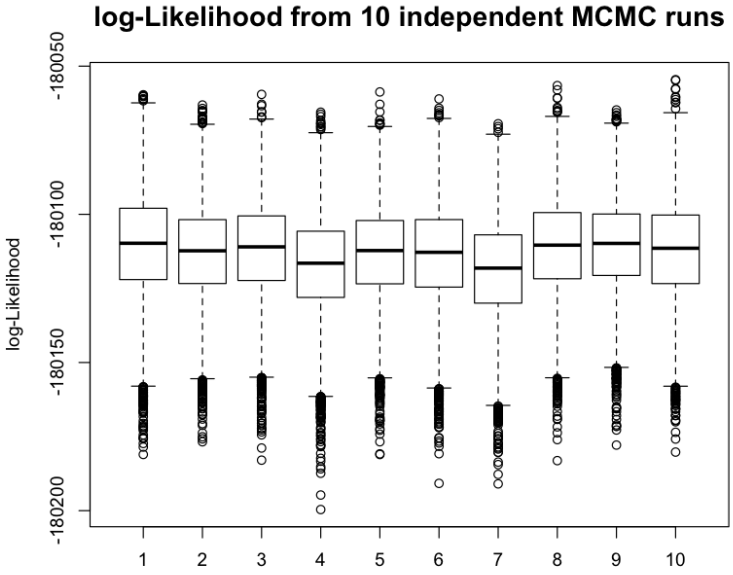


Figure 3.2: Boxplot of log-likelihoods from 10 independent MCMC runs of BPM2 on the same simulated data set. The Gelman and Rubin's potential scale reduction factor (Gelman and Rubin, 1992) on posterior draws of log-likelihood is 1.015.

We apply four methods, BPM1, BPM2, SR and PCA, to 100 simulated data sets. To run BPM1, we need to specify the number of modules and we give BPM1 some advantage

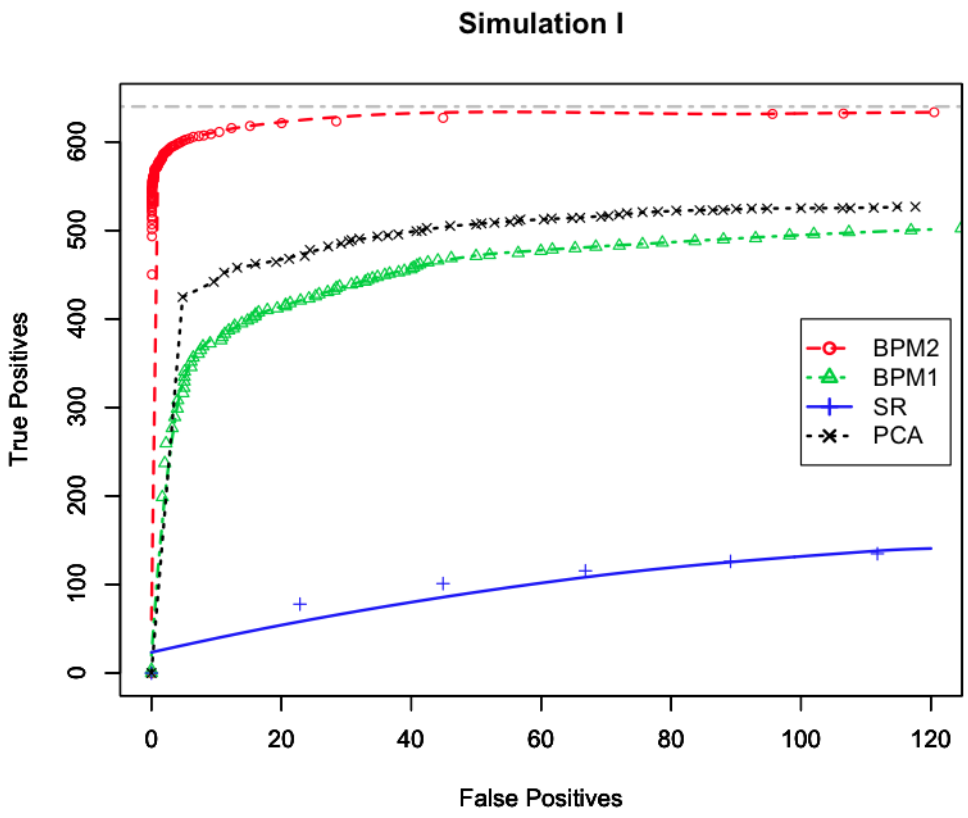


Figure 3.3: The aggregated ROC curves that compare true positives, the total number of the true gene-marker pairs detected, and false positives, the total number of unrelated gene-marker pairs falsely selected, of different methods under simulation in Section 3.5.1. BPM1: the original Bayesian partition model (Zhang et al., 2010); BPM2: the augmented Bayesian partition model proposed in this thesis; SR: a two-stage stepwise method on the one-gene-one-marker regression model (Storey et al., 2005); PCA: a two-stage stepwise method based on the principle component analysis of true genes in each module (oracle benchmark for SR).

by using the true number, $D = 8$. For the augmented BP model, BPM2, we assume that we do not know the true number of gene clusters or modules and use a larger number of gene clusters, $K = 20$. The number of modules is determined by the procedure described in Section 3.4. Figure 3.2 shows the log-likelihoods from 10 independent MCMC runs of BPM2 on the same simulated data set, which demonstrates that the Markov chain used in our method attained a stationary distribution after the burn-in period. The Gelman and

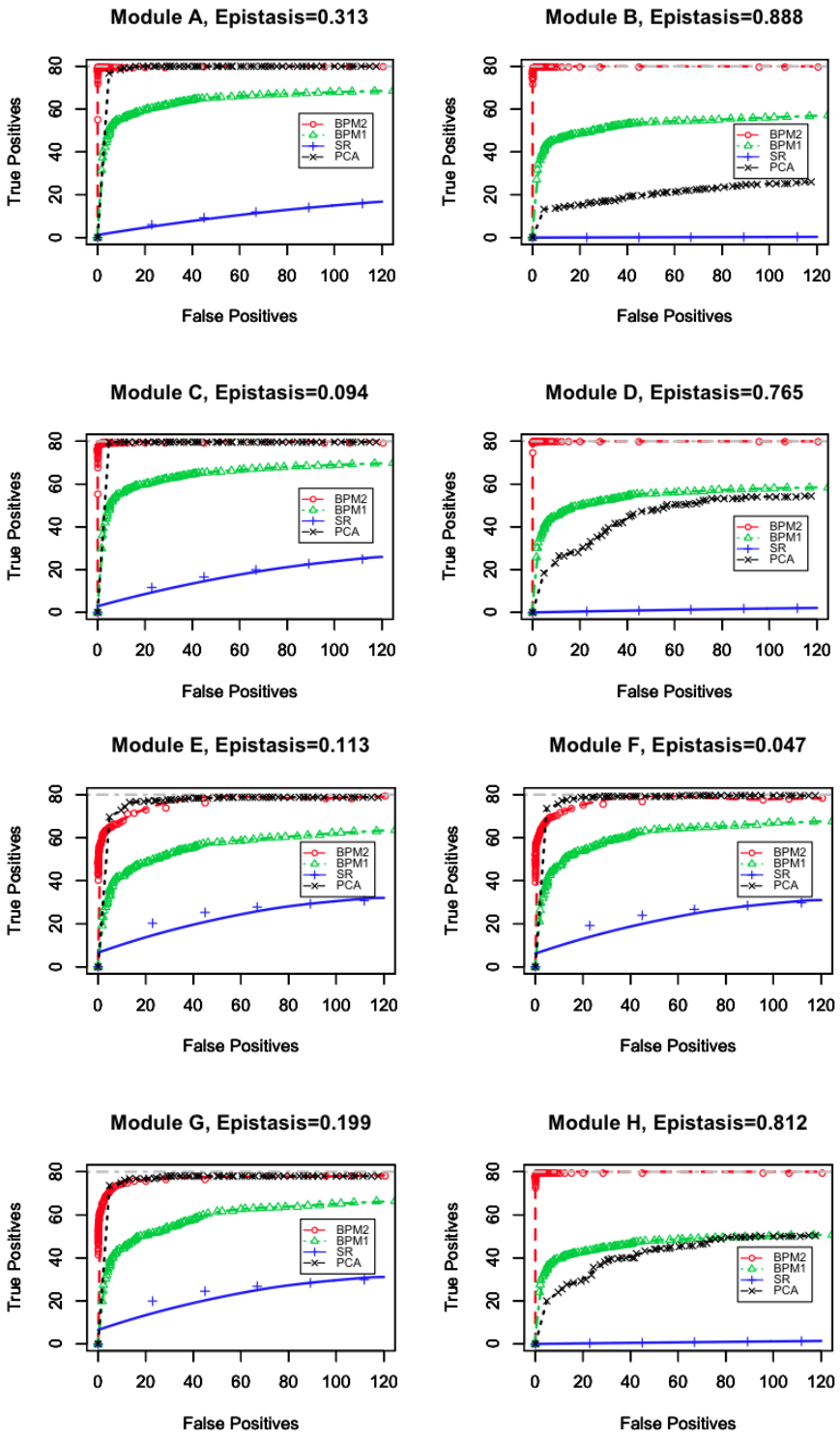


Figure 3.4: The aggregated ROC curves that compare true positives and false positives of different methods in each module under simulation in Section 3.5.1. The average percentage of genetic variance explained by epistasis in a module is shown on top of each figure.

Rubin's potential scale reduction factor (Gelman and Rubin, 1992) based on posterior draws of log-likelihood is 1.015.

The receiver operating characteristic (ROC) curves in Figure 3.3 compare true positives, the total number of the true gene-marker pairs detected, and false positives, the total number of unrelated gene-marker pairs falsely selected, at varying thresholds. Figure 3.4 further compares true positives and false positives of different methods in each module. As shown from the ROC curves in Figure 3.3 and Figure 3.4, in modules that have strong marginal but weak interactive effects, the BPM2 performs almost as well as the PCA method based on the stepwise regression, even though the latter has already been given the set of genes in each module to start with. In modules that have weak marginal but strong interactive effects (module B, D and H), the BPM2 is more powerful than the PCA method in detecting epistatic effects. When the true genes in modules are not given, the stepwise method SR based on the one-gene-one-marker regression model has the lowest detection rate, especially when there are strong epistatic effects. Moreover, the augmented model, BPM2 achieves consistently higher power in detecting eQTLs (gene-marker pairs) compared to the original model, BPM1. There are several reasons for the excellent performance of BPM2. First, BPM2 uses a more efficient algorithm to partition individuals, and a more flexible model of the dependence structure between gene expression and genetic marker. Second, we aggregate information from all the co-regulated genes in a module and improve the signal strength of eQTLs. Third, by using a joint model of interactive markers and a iteratively sampling approach, we significantly increase the power in detecting markers with weak marginal but strong interactive effects compared to the stepwise methods that select one marker at a time.

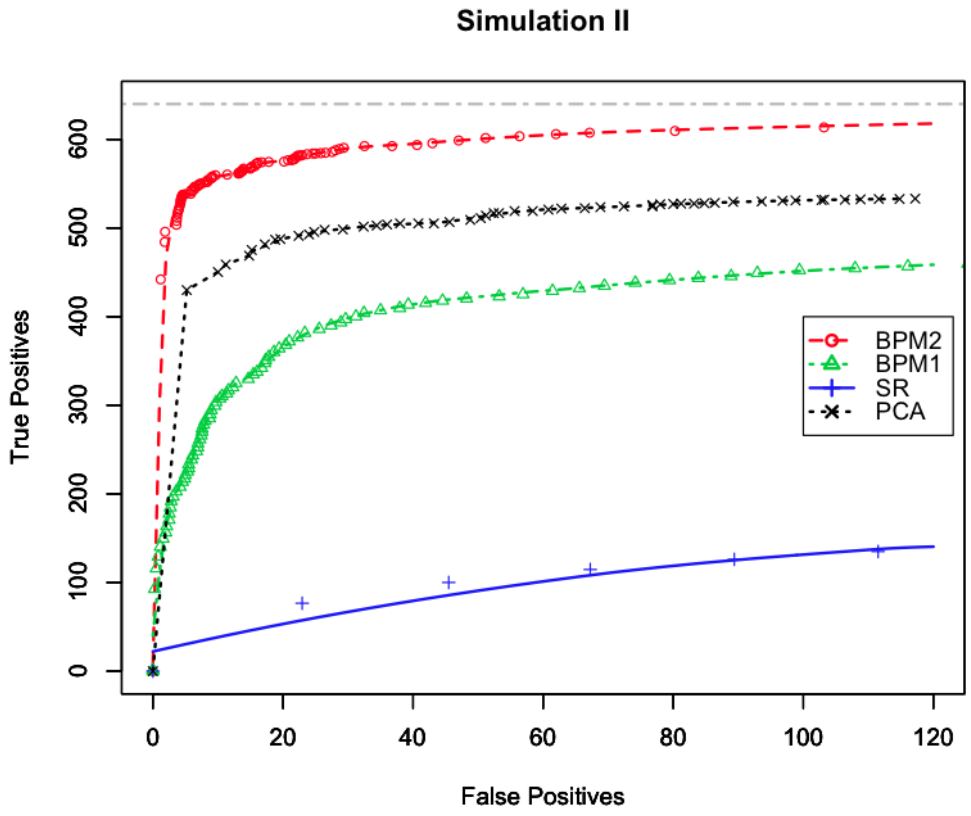


Figure 3.5: The aggregated ROC curves that compare true positives, the total number of the true gene-marker pairs detected, and false positives, the total number of unrelated gene-marker pairs falsely selected, of different methods under simulation in Section 3.5.2.

### 3.5.2 Simulation with Mixed Correlations

The second simulation studies the performance of different methods when there are both positively and negatively correlated genes in the same module. The data generation process is the same as the previous simulation except that a random sign is multiplied to the simulated expression of each gene. Since the original Bayesian partition model cannot capture negatively correlated genes in the same module, we use 16 instead of 8 as the number of modules in BPM1. For the augmented Bayesian partition model, BPM2, we again specify the number of gene clusters as 20 and the number of modules is determined as described in

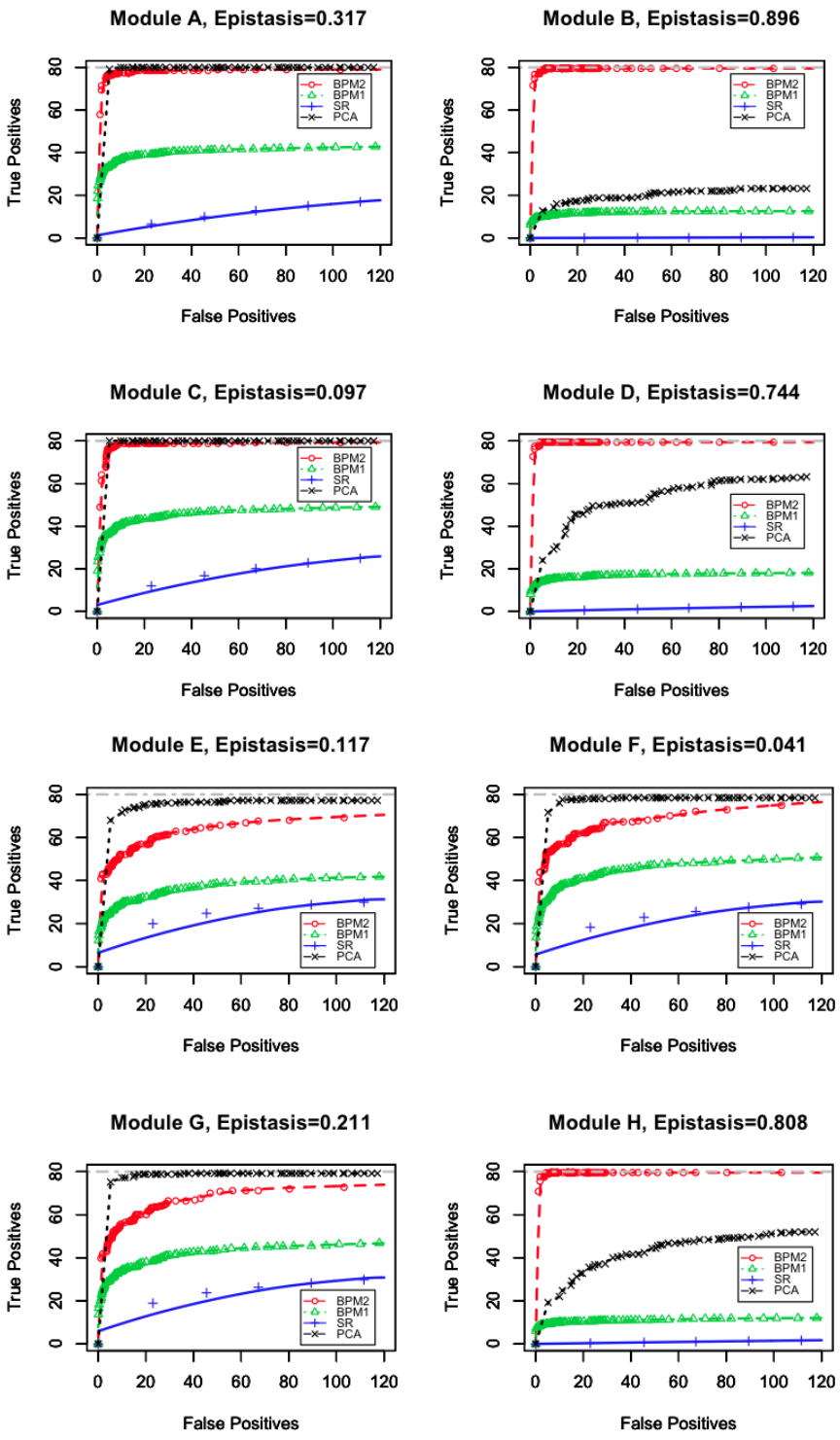


Figure 3.6: The aggregated ROC curves that compare true positives and false positives of different methods in each module under simulation in Section 3.5.2. The average percentage of genetic variance explained by epistasis in a module is shown on top of each figure.

Section 3.4. The aggregated ROC curves of different methods are shown in Figure 3.5 and the ROC curves in each module are shown in Figure 3.6.

As expected, the original Bayesian partition model, BPM1, has a lower power in the second simulation compared to its performance in the first simulation. Although we increased the number of modules in BPM1 from 8 to 16 in order to capture all the relevant genes, the separation of negatively correlated genes into different modules (a module only contains positively correlated genes in BPM1) weakened the signal strength of gene expression in determining individual type partitions. The lower detection rate of BPM1 is more evident in Module B, D and H, when an informative partition of individuals becomes critical in detecting genetic markers with weak marginal but strong interactive effects. To the contrary, the augmented Bayesian partition model, BPM2, is able to combine negatively correlated genes in the same module and shows consistently excellent performances in both simulations. In modules E, F, and G where the major marker explains more than 70% of the genetic variation, the PCA method, which starts with the true gene-module assignments and uses stepwise regression to detect markers, outperforms the BPM2. In Module A and C where the marginal effects of the two marker are almost the same, the BPM2 and the PCA method have comparable performances. In Module B, D and H where no or very weak marginal effect is present and genetic variation is mainly explained by the epistasis, the BPM2 achieved significantly better power than the PCA method, even though the latter has a full knowledge of genes in each module. When the gene-module assignments are not given, the stepwise regression method SR failed to detect most of gene-marker pairs.

## 3.6 Yeast Data Analysis

In this section, we present an application of the augmented Bayesian partition model to a yeast data set with 2957 markers and 3662 gene expression profiles from 112 yeast (*S. cerevisiae*) segregants (Brem and Kruglyak, 2005; Zhang et al., 2010). We set the number of gene clusters $K = 200$ and the number of modules $D = 100$ in this study. Because markers in the yeast data set are very densely distributed, adjacent markers are highly correlated. After MCMC sampling, markers adjacent to the truly linked marker often diluted the posterior probability for the true marker-module linkage. To counter this problem, we first specified a window centered at each marker so that markers inside the window are in high LD with the marker in the center. The posterior probabilities of all markers in the window were summed up and regarded as the modified posterior probability of the central marker. The markers with peak probabilities exceeding the given threshold were selected and all other markers in the corresponding windows were masked out. We chose the window size to contain 5 markers and 0.8 as the threshold for modified posterior probabilities to determine the module membership of a marker. Among 100 modules, we found 40 modules are not associated with any marker above the threshold, 49 modules are associated with a single markers, 10 modules are associated with two markers and 1 module is associated with three markers.

Figure 3.7 shows an example of a module linked to a single marker on Chromosome XII. The genes in the module are grouped into two positively correlated gene clusters with negative correlations between two clusters. The functional annotation of each gene cluster is shown on top of the figure. Out of the 14 genes in the module, nine of them are physically located adjacent to the genetic marker and are cis-acting eQTLs. The other five genes are

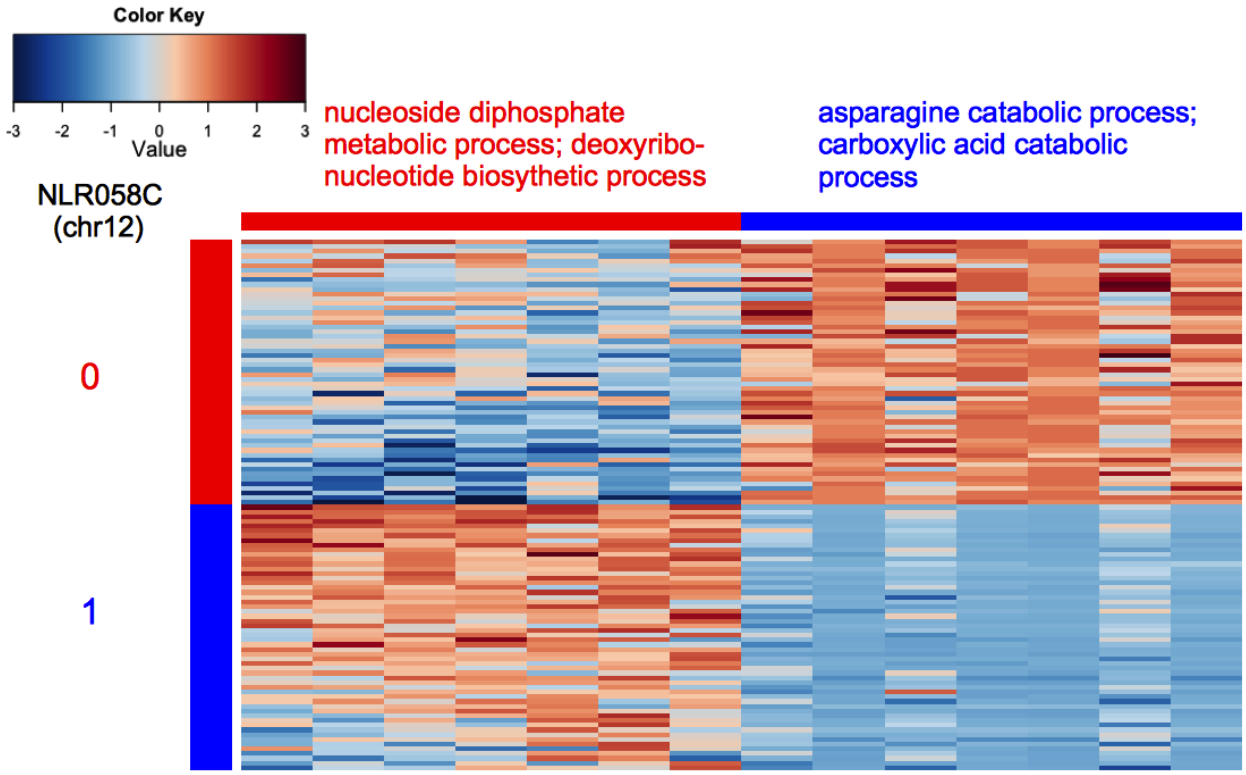


Figure 3.7: Heatmap for gene expression in a module linked to a single marker (NLR058C) on Chromosome XII. Individuals are divided into two groups according to the genotype (0 or 1) of the genetic marker. Each column represents the expression level of a gene across individuals. High- and low-expression levels are represented by red and blue, respectively.

located on different chromosomes and are trans-acting eQTLs.

Figure 3.8 shows a module that are linked to two genetic markers. There are two gene clusters in the module with a total of 27 genes, most of which are related to the sexual reproduction process in yeast. Nine out of 27 genes are located near the marker YCR041W on Chromosome III. The other 18 genes are not located in adjacent to either marker. Box-plots of average gene expression in two gene clusters under different genotype combinations are shown in Figure 3.9. From Figure 3.8 and Figure 3.9, we can see that the marker YCR041W has a primary regulatory effect and divides individuals into two separate groups in both gene clusters. The secondary marker YHL007C further divides the low-expression

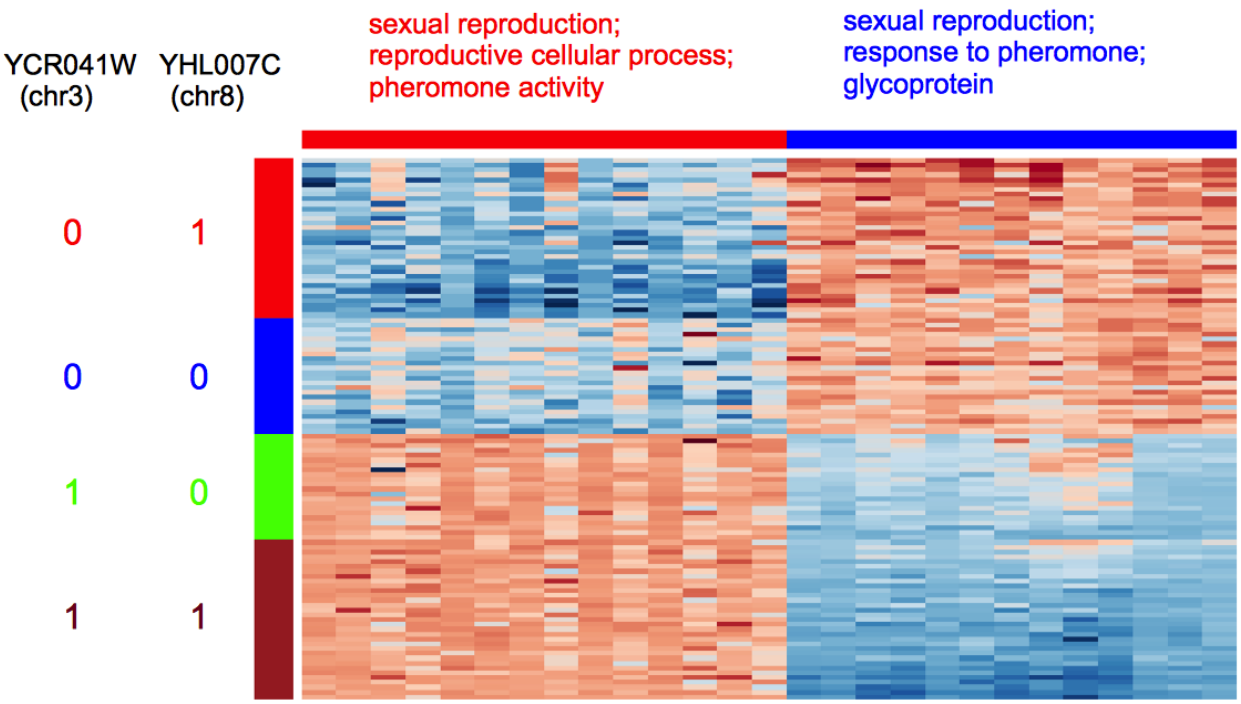


Figure 3.8: Heatmap for gene expression in a module linked to two markers on Chromosome III and VIII. Individuals are divided into four groups according to the genotype combinations of the two markers. Each column represents the expression level of a gene across individuals. High- and low-expression levels are represented by red and blue, respectively.

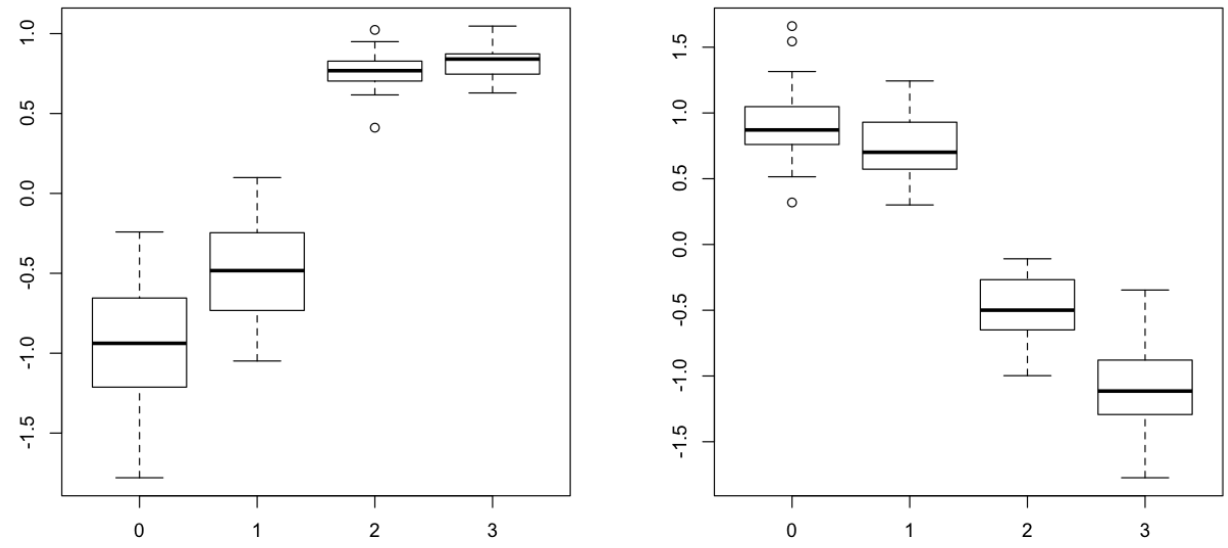


Figure 3.9: Box-plots of average gene expression under different genotype combinations from two gene clusters in Figure 3.8.

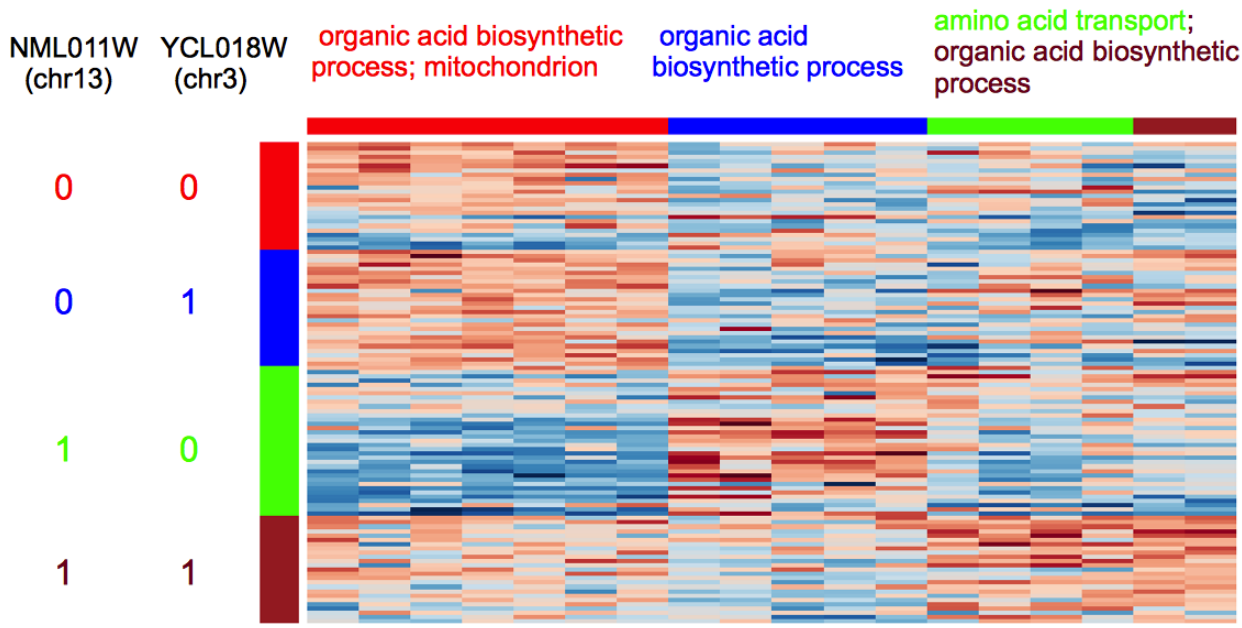


Figure 3.10: Heatmap for gene expression in a module linked to two markers. Individuals are divided into eight groups according to the genotype combinations of the two markers. Each column represents the expression level of a gene across individuals. High- and low-expression levels are represented by red and blue, respectively.

individuals into two subgroups.

Figure 3.10 shows another example of a module that is linked to two genetic markers. The three gene clusters in the module exhibit more complicated gene expression patterns and all of them are involved in organic acid biosynthetic process. Both genetic markers are trans-eQTLs. Individuals with genotype combination $(1, 0)$ from two markers have low expression in the first gene cluster and high expression in the second gene cluster, and individuals with genotype combination $(1, 1)$ have relatively high expression in the third gene cluster.

In the example shown in Figure 3.11, we identified a module linked to three genetic markers. The module consists of four genes with functions related to oxidation-reduction and dehydrogenase. The three genetic markers in the module are trans-eQTLs co-localized with other genes involved in oxidation-reduction, dehydrogenase and ATP-binding respectively.

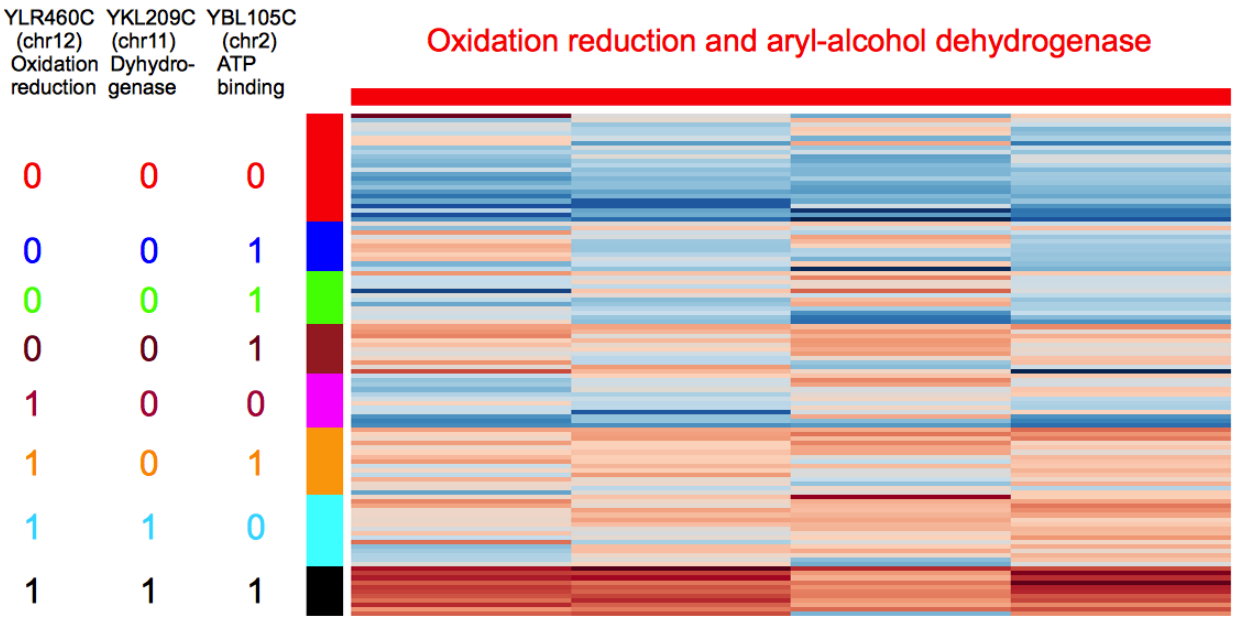


Figure 3.11: Heatmap for gene expression in a module linked to three markers. Individuals are divided into eight groups according to the genotype combinations of the two markers. Each column represents the expression level of a gene across individuals. High- and low-expression levels are represented by red and blue, respectively.

From the heatmap in Figure 3.11, we can see that when the three genetic markers have genotype combination $(1, 1, 1)$, the four trans-acting genes in the module will have relatively higher expression compared with individuals with other genotype combinations.

## 3.7 Discussion

This chapter studied the problem of identifying pleiotropic and epistatic effects in eQTL studies. The contributions of the augmented Bayesian partition model are threefold. First, it improves signal strength by aggregating information from correlated gene clusters and allowing negatively correlated genes to be included in the same module. Second, it directly accounts for dependence structure of genetic markers by modeling them as linkage disequilibrium (LD) blocks. Third, the sequential partition prior incorporates scientifically motivated

assumptions and the corresponding dynamic programming algorithm allows for more efficient computations. Simulation studies have demonstrated that the augmented model achieved significantly improved power in detecting eQTLs compared to the original BP model and traditional regression-based methods. We applied the augmented Bayesian partition model to analyze yeast eQTL data. A particular strength of our method is its ability to detect epistatic effects with high power when the marginal effects are weak, addressing a key weakness of other eQTL mapping methods.

Further improvements of the model are being investigated. First, Zhang (2012) proposed a refined model of the interactions between genetic markers using Bayes networks, which can be incorporated into our Bayesian partition model. Second, using gene expression data from multiple tissues, the Bayesian partition model can be extended to study tissue common and tissue specific eQTLs. We are currently collaborating with scientists in the Genotype-Tissue Expression (GTEx) project, which aims to comprehensively survey genetic regulation of gene expression in multiple human tissues.

# Chapter 4

# Identifying TF Subclasses on Protein Binding Microarray

Transcription factors (TFs) play a key role in the regulation of gene expression by activating or repressing transcription of their target genes. Regulatory specificity is achieved primarily by the recognition of specific DNA binding sites in the genome by sequence-specific TFs. Data on TF DNA binding specificity are important for understanding how transcriptional regulation is encoded in *cis* regulatory sequences in the genome.

TFs can be classified according to the structural class of their DNA binding domains (Luscombe et al., 2000). TFs of the same structural class adopt the same fold in their DNA binding domain and dock with their DNA binding sites in a similar manner. Because of these structural similarities, combined with sequence similarities due to the origin of TF families from ancient gene duplications and subsequent mutations, members of the same DBD class often, but not always, have similar DNA binding sequence preferences (Badis et al., 2009). A DBD class can be further divided into subclasses, with more closely related

proteins exhibiting more similar DNA binding preferences. Understanding how highly similar members of a TF family attain both redundant and divergent regulatory functions remains a significant challenge (Grove et al., 2009).

## 4.1 Background on Protein Binding Microarray

Accurate and comprehensive data on DNA binding sequence specificities are essential for investigations of regulatory targeting by TFs, including the identification of the molecular determinants of TF DNA binding specificity. A variety of high-throughput technologies have been developed for determining TF DNA binding specificity. Methods that provide data on *in vivo* TF occupancies in the genome, such as chromatin immunoprecipitation coupled with DNA microarrays (ChIP-chip) or high-throughput sequencing (ChIP-Seq), provide data on both direct and indirect DNA binding by TFs, which can vary across cellular or environmental conditions (Harbison et al., 2004). In contrast, approaches that determine DNA binding specificities *in vitro* provide data on direct TF-DNA interactions, without the confounding effects of *in vivo* protein cofactors (Gordân et al., 2009).

Protein binding microarray (PBM) technology is an *in vitro* approach for characterizing the DNA binding specificities of proteins, by assaying the binding of a protein to a library of double-stranded DNA sequences immobilized on a DNA microarray (Bulyk et al., 2001). Universal PBMs contain synthetic DNA sequences designed to represent all possible $k$-mers, with commonly used array designs encompassing all possible 10-bp DNA sequence ($k = 10$) (Berger et al., 2006). Universal PBMs have been used in numerous recent studies to determine the DNA binding specificities of hundreds of TFs from a wide range of organisms (Grove et al., 2009; Campbell et al., 2010; Busser et al., 2012), with major efforts on TFs

encoded in the genomes of the yeast *S. cerevisiae* (Badis et al., 2008; Zhu et al., 2009; Gordân et al., 2011) and mouse (Berger et al., 2008; Badis et al., 2009; Wei et al., 2010).

Universal PBMs contain 60-bp DNA probes, each of which contains multiple 10-mers embedded within variable flanking sequence. For statistical robustness, binding preference scores are calculated for all 8-mers, each of which is represented on at least 16 or 32 spots, for palindromic and non-palindromic 8-mers, respectively, on the array. Analysis of universal PBM data using the Universal PBM Analysis Suite including the Seed-and-Wobble algorithm (Berger et al., 2006; Berger and Bulyk, 2009), which were developed together with the universal PBM technology, involves background subtraction, various normalizations of the data, and calculation of various binding scores for each 8-mer, including the median fluorescence signal intensity over all probes that contain a particular 8-mer and a rank-based PBM enrichment (E)score, ranging from $-0.5$ (worst) to $+0.5$ (best). The 8-mer data can be used to derive a DNA binding specificity motif, or position weight matrix (PWM) (Berger et al., 2006; Berger and Bulyk, 2009).

Analyses of large collections of universal PBM data have identified previously unknown diversity in the DNA binding sequences recognized by TFs (Berger et al., 2008; Badis et al., 2009; Gordân et al., 2011). Hierarchical clustering of TFs according to their similarity in 8-mer E-scores has permitted more precise identification of TF subclasses according to their DNA binding specificities (Berger et al., 2008; Wei et al., 2010; Gordân et al., 2011). In addition, examination of $k$-mer binding preferences within a TF family (here, defined as a group of closely related TFs belong to the same DBD structural class or the same subclass within a DBD class) has revealed sets of $k$-mers bound in common across the family (TF-common $k$-mers) and also sets of $k$-mers preferred by an individual member(s) of a TF family. To date, identification of such TF-preferred $k$-mers has been performed in an *ad*

*hoc* fashion, in manual investigations of individual sets of TFs that employed various semi-arbitrary thresholds (Busser et al., 2012) combined with visual inspection (Berger et al., 2008). Such TF-preferred $k$-mers may contribute to the distinct regulatory functions that distinguish members of a TF family.

In the remaining part of this section, we will describe the PBM and ChIP-chip data sets used in our analysis. In Section 4.2, we present a Bayesian hierarchical analysis of variance (ANOVA) approach for modeling PBM $k$-mer data (here, 8-mers) given TF family classifications. Our method identifies 8-mers that score artifactually highly ('sticky' 8-mers) for unknown reasons. Our approach for subsequently adjusting for these systematic biases improves overall PBM data quality and improves concordance with ChIP-chip data. The Bayesian ANOVA model assumes that TFs have been classified into families. In practice, DBD structural class can be used to define TF family memberships. However, members of the same DBD class do not always exhibit similar DNA binding preferences. In Section 4.3, we present a Bayesian partition model for systematically identifying TF subclasses, simultaneously with their shared DNA binding preferences as well as the sequence preferences that distinguish them. Our TF subclassification results are consistent with classifications based on TF DBD sequence similarity. Our method also permits automated identification of TF-preferred $k$-mers within TF subclasses. Improved identification of TF-preferred $k$-mers will aid in studies of potential differences in the targeting of different genomic sites by paralogous TFs, and thus potentially how they may exert different regulatory functions. We anticipate that such modeling will aid in identification of genomic *cis* regulatory codes (*i.e.*, *cis* regulatory sequence features that confer particular gene expression patterns) and will improve the quality of datasets for identification of the molecular determinants of TF-DNA sequence specificity.

### 4.1.1 PBM Data Sets

We downloaded universal PBM $k$-mer data and DBD structural class data from the UniPROBE database (Robasky and Bulyk, 2011), which hosts data generated by universal PBM technology on the *in vitro* DNA-binding specificities of proteins. The relative binding preference of a TF for each $k$-mer (here, $k = 8$) in universal PBMs is quantified by the PBM enrichment score (E-score), which is a modified form of the Wilcoxon-Mann Whitney statistic (Berger et al., 2006). We refer to this as the observed E-score. We consider observed E-scores above 0.35 as corresponding, in general, to sequence-specific DNA binding of the TF. In this study, we included 349 TFs from 19 DBD structural classes (e.g., homeodomain), with the criterion that there are at least three TFs per DBD class. Two of the downloaded PBM data sets are of particular interest in this study. One is a mouse TF data set with 87 TFs from 12 DBD classes (filtered according to the above criterion from a total of 104 TFs from 22 structural classes) previously described by Badis et al. (2009), in which PBM experiments were performed for each TF on two different versions of "all 10-mer" universal arrays, referred to as "version 1" and "version 2" (Agilent Technologies, Inc.; AMADID #015681 (Berger et al., 2008) and #016060 (Zhu et al., 2009), respectively), which were based on two different "all 10-mer" de Bruijn sequences. The other is a yeast TF data set with 79 TFs from 10 DBD classes (filtered according to the above criterion from a total of 89 TFs from 18 structural classes) in Zhu et al. (2009) for which ChIP-chip data (see Section 4.1.2) are publicly available in Harbison et al. (2004) for 57 of these 79 TFs. We also included eight negative control experiments, corresponding to duplicate PBM experiments on each of array design versions 1 and 2, for GST in binding buffer and, separately, for a mock *in vitro* transcription and translation reaction (Badis et al., 2009).

### 4.1.2 ChIP-chip Data Sets

We downloaded yeast ChIP-chip data from Harbison et al. (2004) for 352 ChIP-chip experiments for 207 TFs under various culture conditions. We use the notation *TF_condition* to refer to the ChIP-chip experiment for a transcription factor TF under an environmental condition. For each ChIP-chip data set, we defined the 'bound' intergenic regions to be those with ChIP-chip $P$-value $< 0.001$ and the 'unbound' intergenic regions to be those with ChIP-chip $P$-value $> 0.5$, as reported by Harbison et al. For 57 of these 207 TFs, PBM data are available in UniPROBE. We further restricted our analysis to ChIP experiments for which the ChIP 'bound' regions have been explained as being due to direct DNA binding by the profiled TF (Gordân et al., 2009); this requirement resulted in a final collection of 75 ChIP-chip data sets for 46 TFs.

## 4.2 Bayesian ANOVA Model of PBM $k$-mer data

Given family membership (e.g., DBD structural class or subclass of a DBD class) of TFs, the DNA binding specificity scores from PBM experiments can be decomposed into components attributable to at least three sources of variation: systematic biases across all PBM experiments, family-wise binding preferences shared by members of the same TF family (*i.e.*, TF-common $k$-mers) (Berger et al., 2008; Busser et al., 2012), and $k$-mer binding preferences specific to individual member(s) of a given TF family (*i.e.*, TF-preferred $k$-mers) (Berger et al., 2008; Busser et al., 2012). For PBM data sets from several diverse DBD classes, we used a Bayesian ANOVA model with hidden indicators to decompose PBM E-scores into different components and to infer the corresponding TF-common and TF-preferred $k$-mers systematically.

### 4.2.1 Bayesian ANOVA model for identifying TF-common and TF-preferred $k$-mers

For a TF family $f$ and a $k$-mer $j$, we use $P_{f,j} = 1$ ($P_{f,j} = -1$) to indicate that the $k$-mer is preferred (disfavored) by members of that TF family, and $P_{f,j} = 0$ if members of the family show no consistent preferred or disfavored binding for the $k$-mer. For a TF $i$ and a $k$-mer $j$, we use $Q_{i,j} = 1$ ($Q_{i,j} = -1$) to indicate that the $k$-mer is preferred (disfavored) by the TF, and $Q_{i,j} = 0$ otherwise. Given family membership $F_i = f$ and the *standardized E-score* (standardized to have mean 0 and variance 1) $Y_{i,j}$ of TF $i$ and $k$-mer $j$, we assume the following ANOVA decomposition:

$$Y_{i,j} = \tau_j + \omega_{f,j} + \delta_{i,j} + \epsilon_{i,j}, \tag{4.1}$$

where idiosyncratic noise $\epsilon_{i,j} \sim N(0, \sigma^2)$, systematic background noise $\tau_j \sim N\left(0, \sigma^2/\gamma_1\right)$,

$$\text{family-wise effect } \omega_{f,j} \sim \begin{cases} N\left(\omega_f^+, \sigma^2/\gamma_2\right), & \text{if } P_{f,j} = 1, \\ N\left(\omega_f^-, \sigma^2/\gamma_2\right), & \text{if } P_{f,j} = -1, \\ N\left(0, \sigma^2/\gamma_2\right), & \text{if } P_{f,j} = 0, \end{cases}$$

$$\text{and TF-specific effect } \delta_{i,j} \sim \begin{cases} \delta_i^+, & \text{if } Q_{i,j} = 1, \\ \delta_i^-, & \text{if } Q_{i,j} = -1, \\ 0, & \text{if } Q_{i,j} = 0. \end{cases}$$

We assign inverse chi-square priors on $\sigma^2$, $\gamma_1$ and $\gamma_2$, truncated normal priors on $\omega_f^{\pm}$ and $\delta_i^{\pm}$ (to guarantee model identifiability) and independent multinomial priors on indicators $P$'s and $Q$'s. A graphical representation of the dependence relationship between variables

in Bayesian ANOVA model is given in Figure 4.1. We used a Markov Chain Monte Carlo (MCMC) algorithm (Geman and Geman, 1984; Metropolis et al., 1953) to obtain posterior distribution of parameters and hidden indicators according to the ANOVA model in (4.1) (see Appendix B.2 for detailed description and convergence diagnostics of the MCMC algorithm). In the following study, we are especially interested in the posterior distribution of background noise $\tau_j$, and indicators of family-wise and TF-specific effects.

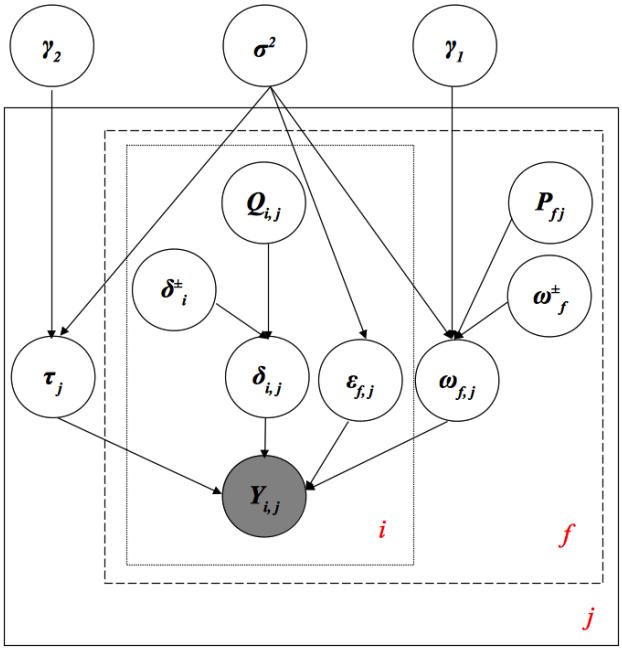


Figure 4.1: A graphical representation of the dependence relationship between the variables in the Bayesian ANOVA model presented in Section 4.2.1.

### 4.2.2 Correcting $k$-mer data for systematic biases

Let $E(\tau_j|\mathbf{Y} = \{y_{i,j}\})$ be the posterior mean of $\tau_j$ calculated from (4.1) given the observed E-scores $\{y_{i,j}\}$, and let $y_j$ be the standardized E-score of $k$-mer $j$ from a PBM experiment. To remove the systematic biases, we subtract the posterior mean of the background noise from the corresponding standardized E-score, *i.e.*, $y^*_j = y_j - E(\tau_j|\mathbf{Y} = \{y_{i,j}\})$. Then, an E-

score corrected for systematic biases can be obtained by transforming $y_j^*$ back to the original scale. We refer to this as the *corrected E-score*.

### 4.2.3 Evaluating the statistical significance of TF-preferred $k$-mers

For a pair of TFs and a given $k$-mer, we evaluate the statistical significance of its being TF-preferred by the intersection-union test (Berger and Hsu, 1996) with the null hypothesis that either none of the TFs exhibits preferred binding to the $k$-mer, or the pair have no difference in their binding preferences for the $k$-mer. Specifically, let $\mu_{i,k}$ be the mean E-score of TF $i$ ($i = 1, 2$) and 6-mer $k$ (there are either 16 or 32 8-mer E-scores corresponding to each 6-mer). For example, the null hypothesis to test if a 6-mer $k$ is preferred by TF 1 is $H_0 : \mu_{1,k} \leq 0$ or $\mu_{1,k} \leq \mu_{2,k}$ and the alternative hypothesis is $H_1 : \mu_{1,k} > 0$ and $\mu_{1,k} > \mu_{2,k}$ (similarly, switch $\mu_{1,k}$ and $\mu_{2,k}$ for testing 6-mers preferred by TF 2). To test $H_0$, we first perform two separate one-side $t$-tests on the averages of $y_{1,j}$ and $d_j = y_{1,j} - y_{2,j}$ for all the 8-mer $j$ that contains the corresponding 6-mer $k$. Then, we can take the maximum of the two $P$-values obtained from $t$-tests as the intersection-union test $P$-value evaluating the significance of TF 1-preferred binding for 6-mer. Finally, we report all TF-preferred $k$-mers at an adjusted $P$-value $< 0.05$ by Benjamini-Hochberg correction (Benjamini and Hochberg, 1995).

## 4.3 Bayesian partition model for identifying TF subclasses

A collection of PBM data sets for TFs from the same DBD class provides a unique perspective to refine the classification of TFs into subclasses according to their DNA bind-

ing sequence preferences. Although hierarchical clustering has been used successfully for functional classification of gene expression profiling microarray experiments (Eisen et al., 1998) and for identification of TF subclasses based on PBM experiments (Berger et al., 2008; Wei et al., 2010; Gordân et al., 2011), the inclusion of unnecessary features that are irrelevant to cluster determination may degrade the results. This is especially the case for PBM experiments, where only a small fraction of $k$-mers measured by experiments are bound specifically by the profiled TF. Biclustering is a simultaneous similarity-based clustering approach that is able to detect subsets of features that exhibit consistent patterns over subsets of experiments (Cheng and Church, 2000; Gusenleitner et al., 2012); however, it does not directly provide a systematic classification of TF subclasses. Model-based methods for identifying and removing batch effects and other sources of variation have been developed for meta-analysis of high-throughput data, including microarray-based gene expression profiling experiments (Johnson et al., 2007; Leek and Storey, 2007, 2008; Leek et al., 2010, 2012). Direct applications of such methods potentially could separate systematic background noise from identification of $k$-mers preferred by different TFs from PBM data.

Here, we present a Bayesian partition model that simultaneously partitions TFs into *subclasses* that have similar DNA binding profiles, and clusters $k$-mer DNA sequences into *groups* that are preferred by one or more TF subclasses.

### 4.3.1 Bayesian partition model of $k$-mers preferred by TF subclasses

The partition model of E-scores for $k$-mers preferred by TF subclasses is very similar to the hierarchical model of expression levels for gene clusters given individual types in

Section 3.3.1. Specifically, let $Y_{i,j}$ be the standardized E-score of TF $i \in \{1, 2, \ldots, N_T\}$ and $k$-mer $j \in \{1, 2, \ldots, N_K\}$, where $N_T$ is the number of PBM data sets for TFs from the same DBD structural class and $N_K$ is the total number of $k$-mers after collapsing forward and reverse complements ($N_K = 32896$ for $k = 8$). Suppose $C_i$ is the unknown subclass of TF $i$ and $G_j$ is the unknown group membership of $k$-mer $j$. For each group $g$ of the $k$-mers ( $g = 1, 2, \ldots, N_G$), $I_g = 1$ if the group is preferred by one or more TF subclasses, and $I_g = 0$ otherwise. Given $C_i = c$ and $G_j = g$, we assume:

$$Y_{i,j}|G_j = g \sim N(\mu_{i,g}, \sigma^2), \text{ and } \mu_{i,g}|C_i = c \sim N\left(\theta_{c,g}, \sigma^2/\kappa_1\right), \tag{4.2}$$

where $\theta_{c,g}$ follows $N\left(0, \sigma^2/\kappa_2\right)$ if $I_g = 1$ and $\theta_{c,g} = 0$ if $I_g = 0$. A graphical representation of the dependence relationship between variables in the Bayesian partition model is given in Figure 4.2. We further assume that $\sigma^2$ follows an inverse chi-square prior Inv-$\chi^2\left(\nu_0, \sigma_0^2\right)$. We found that the results are not sensitive to the choices of hyper-parameters (see Appendix B.3), and we set $\kappa_1 = \kappa_2 = \nu_0 = \sigma_0^2 = 1$. Since the total number of subclasses is unknown, we assume subclass assignments $\mathbf{C} = \{C_i : 1 \leq i \leq N_T\}$ follows a Dirichlet process prior with concentration parameter $\alpha = 0.01$. The prior probability of group assignment $\mathbf{G} = \{G_j : 1 \leq j \leq N_K\}$ is given by $\pi(G_j = g) = 1/N_G$ for $1 \leq g \leq N_G$ and $N_G = 100$, and the prior probability of subclass-preferred indicators $\mathbf{I} = \{I_g : 1 \leq g \leq N_G\}$ is $\pi(I_g = 1) = 0.01$ for $1 \leq g \leq N_G$. Sensitivity analysis on choices of hyper-parameters and priors is discussed in Appendix B.3.

For each group $g$ of the $k$-mers, given $\mathbf{C}$ and $I_g$, we can integrate out (*i.e.*, marginalize over) intermediate parameters in the hierarchical model (4.2) to get an explicit expression of the probability $P\left(\mathbf{Y}_g|I_g, \mathbf{C}\right)$, where $\mathbf{Y}_g = \{Y_{i,j} : G_j = g, 1 \leq i \leq N_T\}$ is the collection of

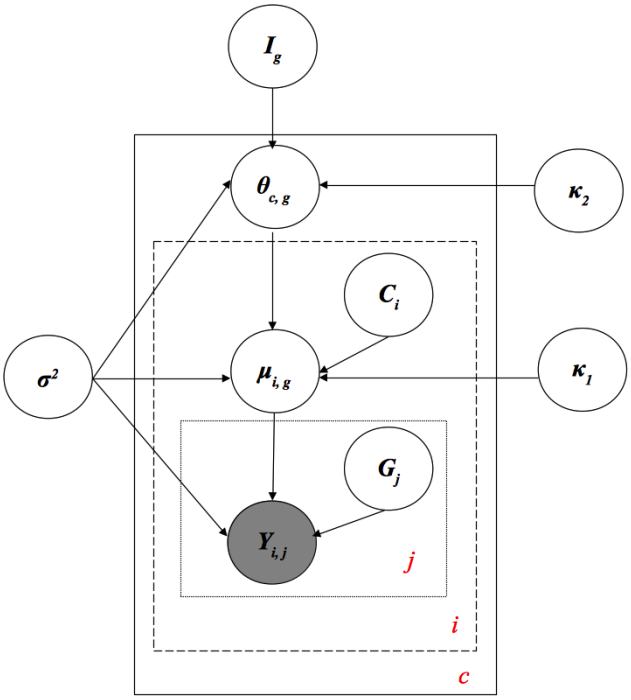


Figure 4.2: A graphical representation of the dependence relationship between the variables in the Bayesian partition model presented in Section 4.3.1.

E-scores for $k$-mers in the group $g$. Let $N_C$ be the number of distinct subclasses, $n_c$ be the number of TFs in subclass $c$ and $n_g$ be the number of $k$-mers in group $g$. After integrating out $\mu_{i,g}$ and $\theta_{c,g}$, we have

$$\Pr\left(\mathbf{Y}_g | I_g = 1, \mathbf{C}, \sigma^2\right) = \left(2\pi\sigma^2\right)^{\frac{N_T n_g}{2}} Z_1 \exp\left(-\frac{S_1^2}{2\sigma^2}\right),$$

where

$$Z_1 = \left(\sqrt{\frac{\kappa_1}{n_g + \kappa_1}}\right)^{N_T} \left(\prod_{c=1}^{N_C} \sqrt{\frac{(n_g + \kappa_1)\kappa_2}{(n_g + \kappa_1)\kappa_2 + n_g n_c \kappa_1}}\right),$$

and

$$S_1^2 = \sum_{i=1}^{N_T} \left[\sum_{j=1}^{n_g} y_{i,j}^2 - \frac{\left(\sum_{j=1}^{n_g} y_{i,j}\right)^2}{n_g + \kappa_1}\right] - \sum_{c=1}^{N_c} \frac{\kappa_1^2 \left(\sum_{C_i=c} \sum_{j=1}^{n_g} y_{i,j}\right)^2}{(n_g + \kappa_1)\left[(n_g + \kappa_1)\kappa_2 + n_g n_c \kappa_1\right]}.$$

We can further integrate out $\sigma^2$ given its inverse chi-square prior Inv-$\chi^2(\nu_0, \sigma_0^2)$,

$$\Pr(\mathbf{Y}_g|I_g=1,\mathbf{C}) = Z_1 \frac{\Gamma\left(\frac{N_T n_g+\nu_0}{2}\right)(\nu_0\sigma_0^2)^{\frac{\nu_0}{2}}}{[\Gamma(1/2)]^{N_T n_g}\,\Gamma(\nu_0/2)\,(S_1^2+\nu_0\sigma_0^2)^{\frac{N_T n_g+\nu_0}{2}}}.$$

Similarly, when $I_g = 0$ (*i.e.*, $\mathbf{Y}_g$ have the same distribution across different subclasses), we have

$$\Pr(\mathbf{Y}_g|I_g=0,\mathbf{C}) = Z_0 \frac{\Gamma\left(\frac{N_T n_g+\nu_0}{2}\right)(\nu_0\sigma_0^2)^{\frac{\nu_0}{2}}}{[\Gamma(1/2)]^{N_T n_g}\,\Gamma(\nu_0/2)\,(S_0^2+\nu_0\sigma_0^2)^{\frac{N_T n_g+\nu_0}{2}}},$$

where

$$Z_0 = \left(\sqrt{\frac{\kappa_1}{n_g+\kappa_1}}\right)^{N_T}\sqrt{\frac{(n_g+\kappa_1)\kappa_2}{(n_g+\kappa_1)\kappa_2+n_g N_T\kappa_1}},$$

and

$$S_0^2 = \sum_{i=1}^{N_T}\left[\sum_{j=1}^{n_g} y_{i,j}^2 - \frac{\left(\sum_{j=1}^{n_g} y_{i,j}\right)^2}{n_g+\kappa_1}\right] - \frac{\kappa_1^2\left(\sum_{i=1}^{N_T}\sum_{j=1}^{n_g} y_{i,j}\right)^2}{(n_g+\kappa_1)\left[(n_g+\kappa_1)\kappa_2+n_g N_T\kappa_1\right]}.$$

Combining with prior distributions of $\pi(\mathbf{C})$, $\pi(\mathbf{G})$ and $\pi(\mathbf{I})$, we obtain the posterior distribution of $\mathbf{C}$, $\mathbf{G}$ and $\mathbf{I}$ given observed E-scores $\mathbf{Y}$,

$$\Pr(\mathbf{C},\mathbf{G},\mathbf{I}|\mathbf{Y}) \propto \pi(\mathbf{C})\pi(\mathbf{G})\pi(\mathbf{I})\prod_{g=1}^{N_G}\Pr(\mathbf{Y}_g|I_g,\mathbf{C}). \tag{4.3}$$

We can draw from the above posterior distribution (4.3) iteratively using a collapsed Gibbs sampler (Liu, 1994) (see Appendix B.3 for detailed description and convergence diagnostics of the Gibbs sampling algorithm).

### 4.3.2 Motif model for aligning $k$-mer DNA sequences

We build a position weight matrix (PWM) to characterize the DNA binding specificity of a group of $k$-mers (e.g., TF-common $k$-mers for a TF family). An element of PWM

$Q = (q_{m,n})$ is defined as the probability of observing a nucleotid $n \in \{\text{A,C,G,T}\}$ at position $m \in \{1, 2, \ldots, W\}$, where $W$ is the pre-determined length of the PWM (here, $W = 10$ for 8-mer PBM data). Let $S_j = \{s_{j,l}, l = 1, 2, \ldots, k\}$ be the DNA sequence of $k$-mer $j \in \{1, 2, \ldots, n_g\}$ in group $g$, where $n_g$ is the number of $k$-mers in group $g$ , and $a_j$ is the alignment position of $k$-mer $j$ within the PWM, where $a_j \in \{-4, -3, \ldots, W - k + 4\}$ . Here, we allow the $k$-mer sequence not to be fully "contained" within a PWM but instead require that the alignment have an overlap of at least 4 nucleotides. For example, $a_j = -2$ means that the third position of $k$-mer $j$ is aligned with the start (*i.e.*, 5' end) of the PWM. The background probability of nucleotide $n \in \{\text{A,C,G,T}\}$, $r_n$, is assumed to be 0.25. The probability of sequence being generated by the motif model is then given by:

$$P\left(S_j | a_j, Q\right) = \prod_{l=1}^{k} \left( r_{s_{j,l}} \mathbb{I}_{\{a_j + l \leq 0\}} + q_{a_j + l, s_{j,l}} \mathbb{I}_{\{1 \leq a_j + l \leq W\}} + r_{s_{j,l}} \mathbb{I}_{\{a_j + l > W\}} \right), \tag{4.4}$$

where $\mathbb{I}_A$ is an indicator function of event $A$. We assign a uniform prior over possible alignment positions $\{-4, -3, \ldots, W-k+4\}$ for $a_j$ $(1 \leq j \leq n_g)$ and a Dirichlet distribution prior Dirichlet$(0.1, 0.1, 0.1, 0.1)$ for each column $(q_{m,n})_{n \in \{\text{A,C,G,T}\}}$ of PWM Q. Then, we can use the same Gibbs sampling strategy as described in (Lawrence et al., 1993) to iteratively update $\{a_j\}_{1 \leq j \leq n_g}$ and Q according to (4.4). Finally, we construct a PWM based on the posterior modes of and generate a corresponding motif sequence logo (Schneider and Stephens, 1990).

## 4.4 Results

We have developed a Bayesian ANOVA model to decompose 8-mer PBM E-scores into background noise (*i.e.*, artifactually high-scoring background $k$-mers), family-wise effects and experiment-specific effects, given a collection of PBM experiments and TF family classifica-

tion based on DBD structural class. Below we start with the identification of artifactually high-scoring background $k$-mers, then describe the identification of TF subclasses based on the $k$-mer data, and conclude with identification of experiment-specific effects in analyses aimed at improved identification of sets of $k$-mers bound preferentially by one TF as compared to a closely related TF (*i.e.*, TF-preferred $k$-mers).

### 4.4.1 Identification of artifactually high-scoring ('sticky') $k$-mers

From the ANOVA model (4.1) described in Section 4.2.1, we can infer $k$-mer background noise based on their posterior means. Background noise constitutes a non-negligible component of E-scores with a standard deviation of 0.063, as compared to a standard deviation of 0.148 for E-scores (Figure 4.3a). The posterior means of the variance and scale parameters $\sigma^2$, $\gamma_1$ and $\gamma_2$ are 0.647, 1.985 and 1.505, respectively. Examination of the sequences of the top 50 artifactually high-scoring $k$-mers, ranked according to their background noise and their E-scores across 357 experiments in our PBM data sets, indicates that AT-rich $k$-mers have artifactually high E-scores in nearly all PBM data sets for a diverse range of TF DBD classes (Figure 4.3c); the most 'sticky' $k$-mer across a wide range of TF DBD classes is AAAAAAAA (Figure 4.3b).

To compare the background noise of $k$-mers on different array designs, we calculated the $k$-mer background noise from the mouse TF PBM data from Badis et al. (2009), in which 2 different de Bruijn sequence array designs were used in PBM experiments for each TF; these are designated as array design versions 1 and 2 (AMADID #015681 (Berger et al., 2008) and #016060 (Zhu et al., 2009), respectively). Comparison of the 'sticky' $k$-mers (*i.e.*, those with background noise larger than one standard deviation, which is 0.148 for version 1 and 0.153 for version 2) from this dataset indicates that (Figure 4.3d) the two different

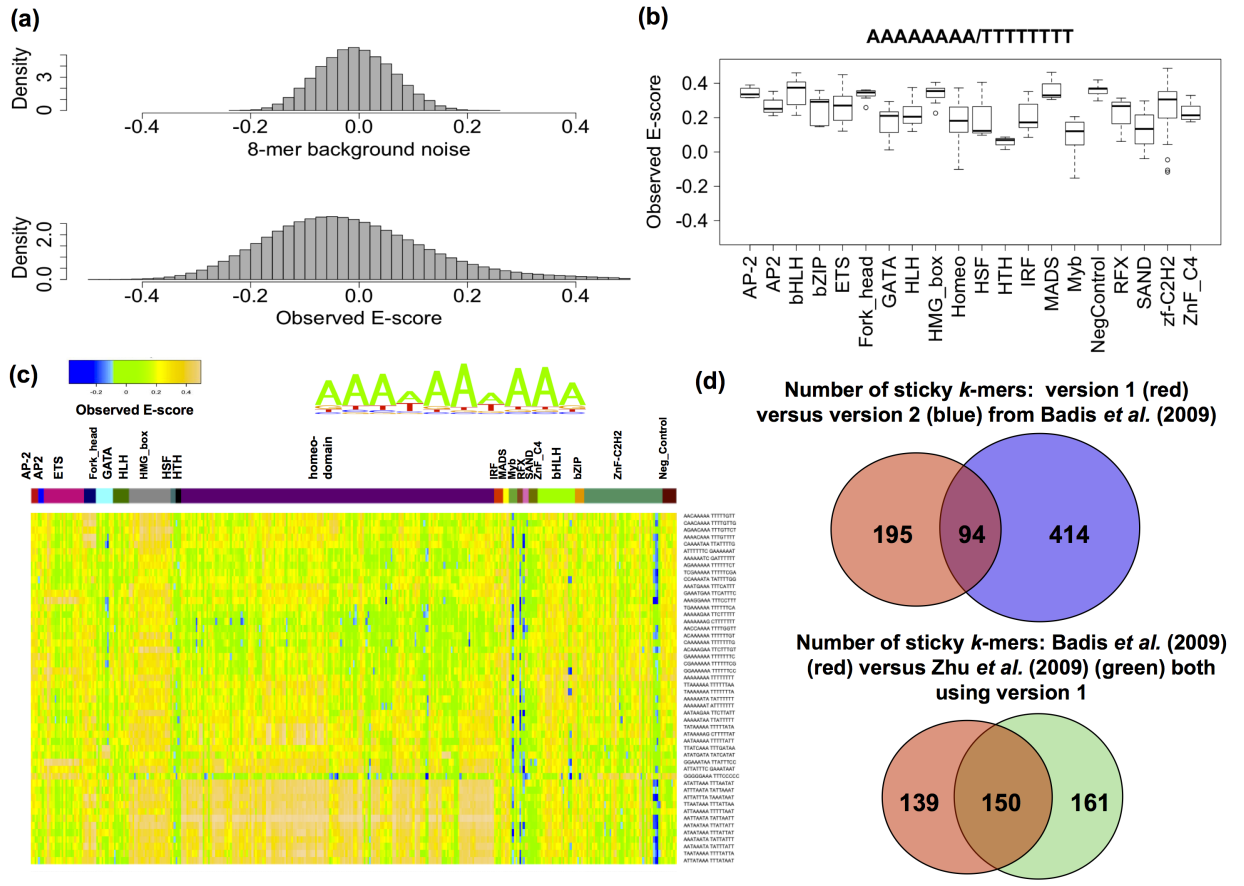


Figure 4.3: Artifactually high-scoring background $k$-mers. (a) Comparison of the distribution of $k$-mer background noise and the original (observed) E-scores. (b) Box plot of E-scores for the most 'sticky' $k$-mer (AAAAAAAA, collapsed with reverse complement TTTTTTTT) across all the available PBM data sets, including a set of negative control PBM experiments . (c) Sequence motif logo generated from the top 50 artifactually high-scoring $k$-mers and their E-scores across our PBM data sets. The multi-colored strip above the heatmap indicates each TF's DBD class (from left to right): AP-2, AP2, ETS, Fork_head, GATA, HLH, HMG_box, HSF_DNA-bind, HTH, Homeodomain, IRF, MADS, Myb, RFX, SAND, ZnF_C4, bHLH, bZIP, Zf-C2H2 and negative control experiments. (d) (Top) Venn diagram comparing the number of 'sticky' $k$-mers with background noise larger than one standard deviation from two different "all 10-mer" de Bruijn sequence array designs; (Bottom) Venn diagram comparing 'sticky' $k$-mers identified from and , both of which used array design version 1.

array versions exhibit different numbers of 'sticky' $k$-mers, and that array design version 2 is noisier (*i.e.*, has a larger number of 'sticky' $k$-mers) than version 1. The two different array designs exhibited some differences in $k$-mer background noise (Pearson correlation coefficient $r = 0.65$; Figure 4.5a), as compared with independent experiments using the same array design (Pearson correlation coefficient $r = 0.88$; Figure 4.5b); for example, although AAAAAAAA is artifactually high scoring in both version 1 and version 2 data sets, CCCCGCCC is found to be 'sticky' only in version 1 data sets (Figure 4.4). To further investigate this effect, we compared the results from the Badis et al. mouse TF PBM data against the set of 'sticky' $k$-mers that we identified in analysis of a separate yeast TF PBM data set (Zhu et al., 2009), both of which used version 1 arrays. We observed significant overlap in the 'sticky' $k$-mers identified in these different datasets (Figure 4.3d); differences in these sets of 'sticky' $k$-mers could be due to differences in protein sample preparation, experimental variation, and differences in the representation of different DBD classes among the TFs that were tested in PBMs in the Badis et al. (2009) versus Zhu et al. (2009) studies.

### 4.4.2 Correction for artifactually high-scoring background for 'sticky' $k$-mers

As described in Section 4.2.2, we correct the E-score of each $k$-mer by subtracting its background noise from the original (observed) E-score. For example, for PBM data for the yeast TF Rpn4 published in Zhu et al. (2009), the observed E-scores and their corresponding background noise are highly correlated (Pearson correlation coefficient $r = 0.71$; Figure 4.6a); subtracting the background noise from the E-scores reduced this correlation to 0.26. Comparison of the motif logo generated from the 68 $k$-mers with observed E-scores greater

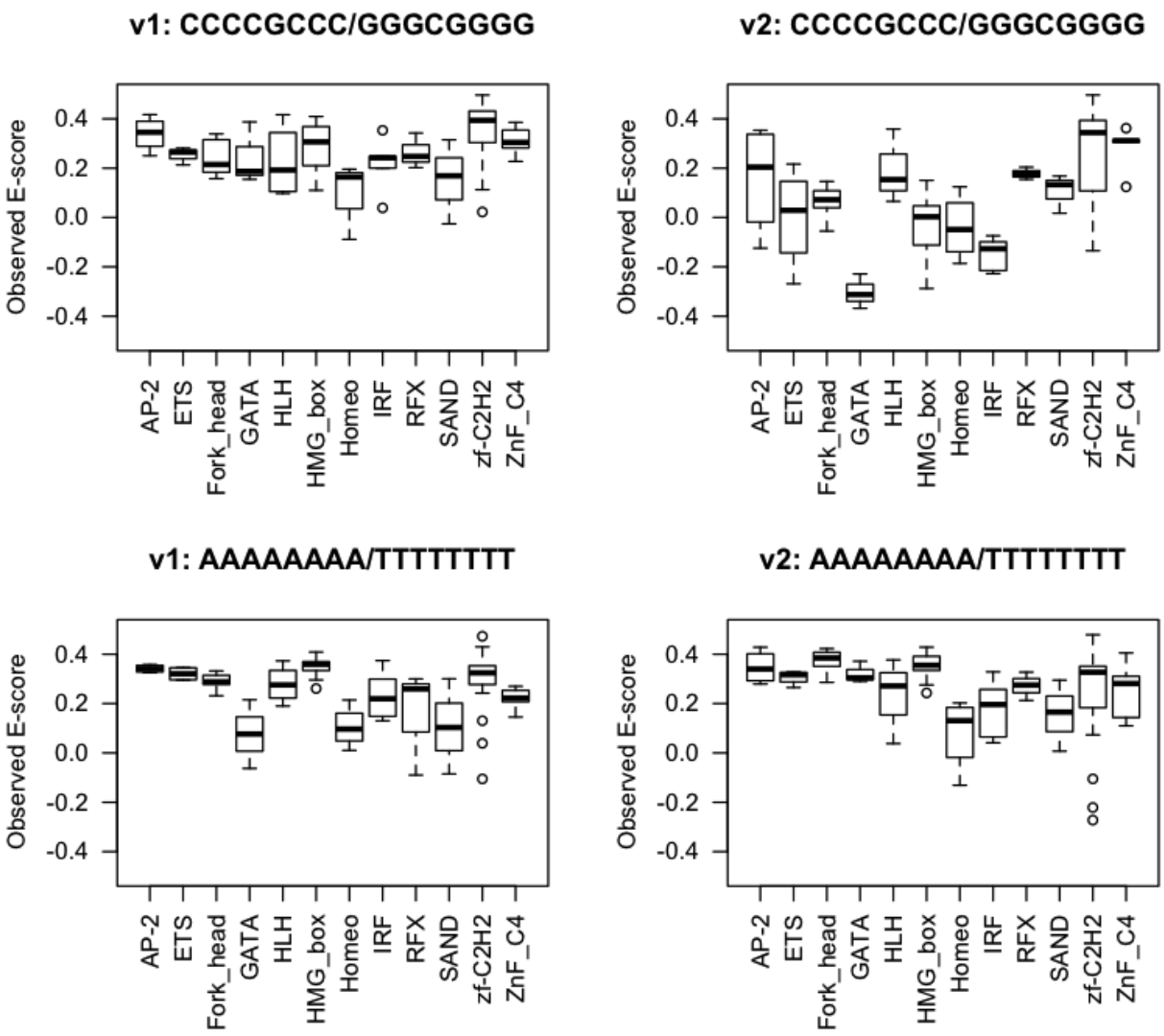


Figure 4.4: Box-plots of observed E-scores for two 'sticky' 8-mers AAAAAAAA and CCCCGCCC in version 1 and their E-scores in version 2.

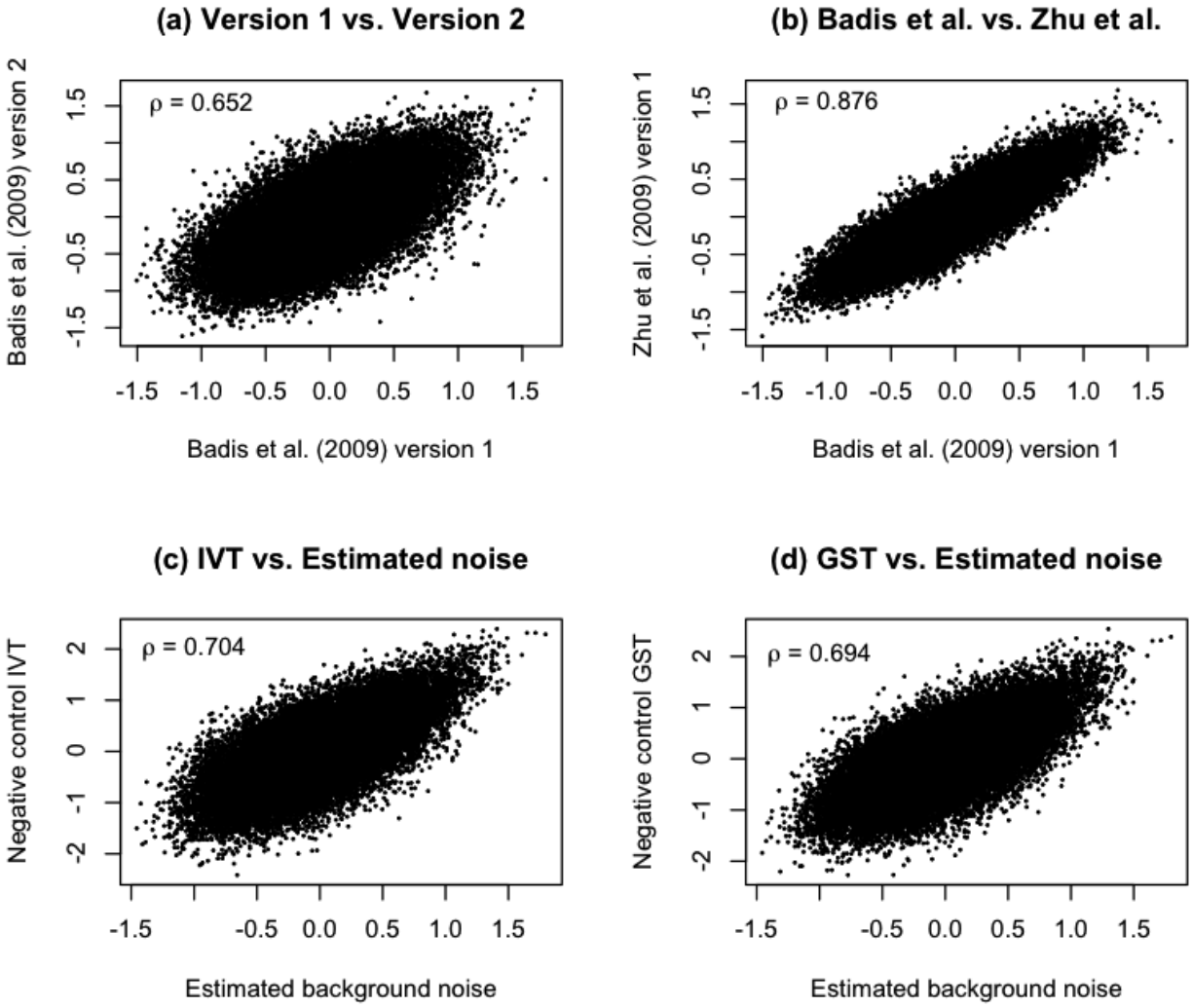


Figure 4.5: (a) Comparison of 8-mer background noise between array design version 1 and version 2 from Badis et al. (2009). (b) Comparison of 8-mer background noise between Badis et al. (2009) and Zhu et al. (2009) both using version 1. (c) Comparison between 8-mer background noise estimated from PBM experiments on 349 TFs and 8-mer E-scores from a mock in vitro transcription and translation reaction (IVT) experiment using version 1. (d) Comparison between 8-mer estimated background noise and 8-mers from an experiment for GST in binding buffer (GST) both using version 1.

than 0.35 versus the logo generated from the same number of top scoring $k$-mers after correction for background noise shows that the quality of the motif greatly improves by correcting the E-scores for systematic biases (Figure 4.6b).

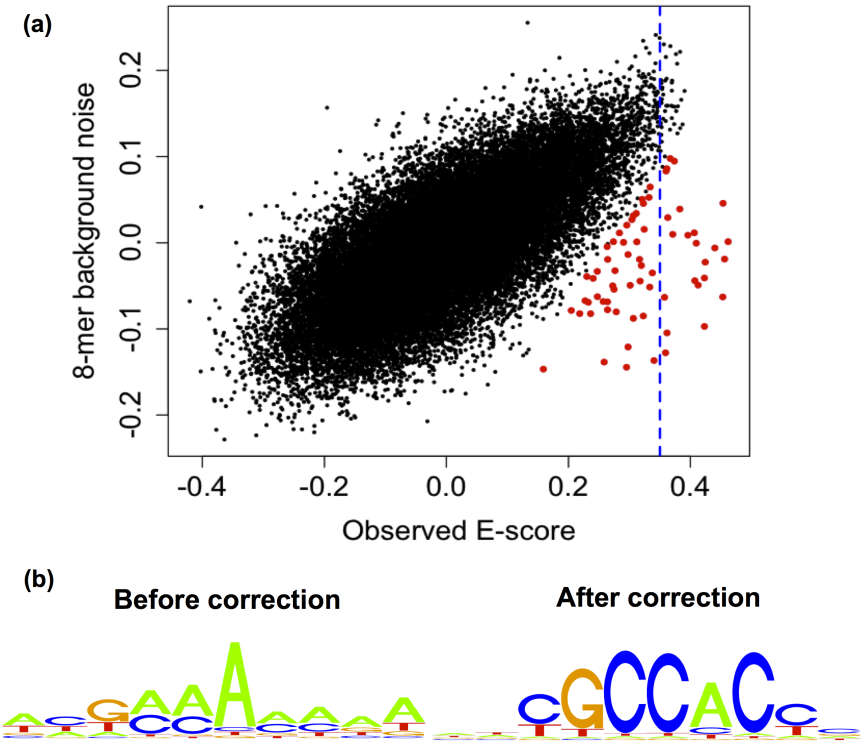


Figure 4.6: Correction for artifactually high-scoring background $k$-mers. (a) Scatter plot of $k$-mer observed E-scores and background noise for yeast TF Rpn4. The blue dotted line indicates the original threshold of E-score $\geq 0.35$. Red points indicate specifically bound 8-mers after background correction. The correlation between E-scores and background noise diminishes from 0.71 to 0.26 after correction. (b) Improvement in motif quality for Rpn4 after correction for $k$-mer background noise.

### 4.4.3 Evaluation of corrected $k$-mer E-scores as compared to ChIP-chip data

We used *in vivo* ChIP-chip binding data to further evaluate the effect of background noise correction of PBM E-scores. Specifically, we first applied background noise correction to the yeast PBM data from Zhu et al. (2009), and then assessed whether this resulted in an improvement in scoring of regions called as 'bound' in the Harbison et al. ChIP-chip data (Harbison et al., 2004) for the same TF. Briefly, for a given TF and a given intergenic sequence, we first calculated an occupancy score by summing PBM median signal intensities for each $k$-mer with an observed E-score above 0.35. We used these PBM-based occupancy scores to rank the intergenic sequences within the ChIP-chip 'bound' and 'unbound' regions for each ChIP-chip data set, and then calculated the corresponding area under the receiver operating characteristic (ROC) curve (AUC statistic). We repeated this same AUC calculation using the corrected E-scores. For comparison, we rank $k$-mers by their corrected E-scores and score the intergenic sequences using the same number of top-ranked $k$-mers as in the calculation with the original (observed) E-scores. In practice, the background correction of a new PBM experiment is based on the estimation from previous experiments; to accurately evaluate this process, we used an independent mouse TF PBM data set (with the same array design) from Badis et al. (2009) to calculate $k$-mer background noise.

Overall, use of the corrected E-scores resulted in a statistically significant increase ($P$-value $= 2.8 \times 10^{-4}$ by Student's $t$-test) in AUC statistics (Figure 4.7). Some ChIP-chip data sets exhibited a sharp increase in AUC by using the corrected E-scores; for example, the AUC for Rpn4 increased from 0.573 to 0.749 for ChIP-chip data for a highly hyperoxic condition (RPN4_H2O2Hi) and from 0.680 to 0.910 for ChIP-chip data for a mildly hyperoxic condition

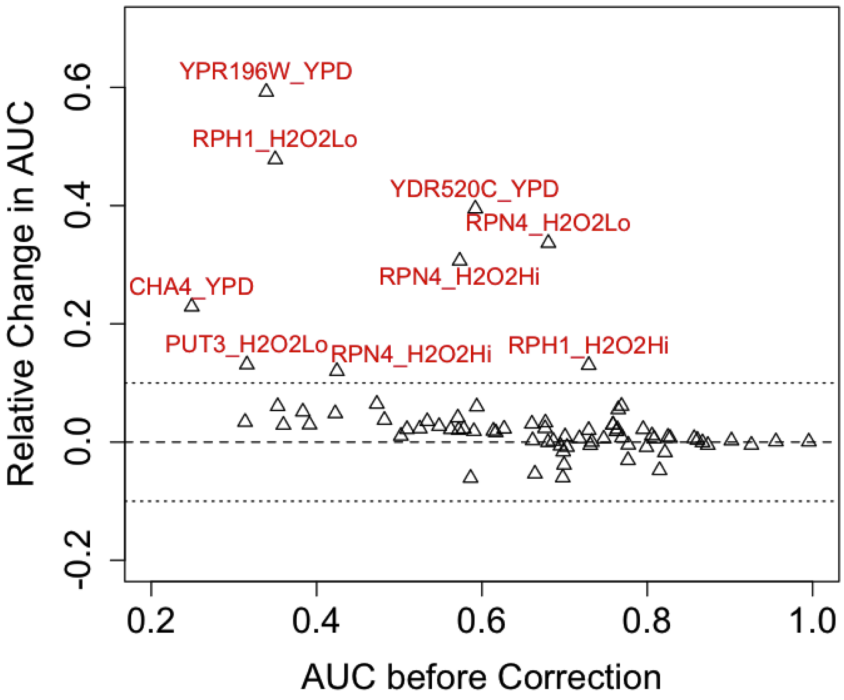


Figure 4.7: Comparison of corrected PBM $k$-mer data to ChIP-chip experiments. The relative change in AUC values by using corrected E-scores is plotted against the original AUC values before background noise correction. ChIP-chip data sets for which the use of corrected E-scores resulted in at least 10.0% change in AUC value are indicated in red with *TF_condition* names.

(RPN4_H2O2Lo). Moreover, PBM data sets with relatively few high-scoring $k$-mers and low AUC values in such ChIP-chip analysis showed a uniform increase in AUC with the use of corrected E-scores; this observation is consistent with the hypothesis that the effect of $k$-mer background noise is more prominent in PBM data sets for TFs with relatively weaker binding signal. Any decreases in AUC value from using corrected E-scores were minor (less than 6.0%), and in one extreme case — Yap6, for which use of corrected E-scores resulted in a decrease of 6.0% from the original AUC value — the difference appears to be due to AT-rich 'sticky' $k$-mers that are bound sequence-specifically by certain TFs, such as Sum1,

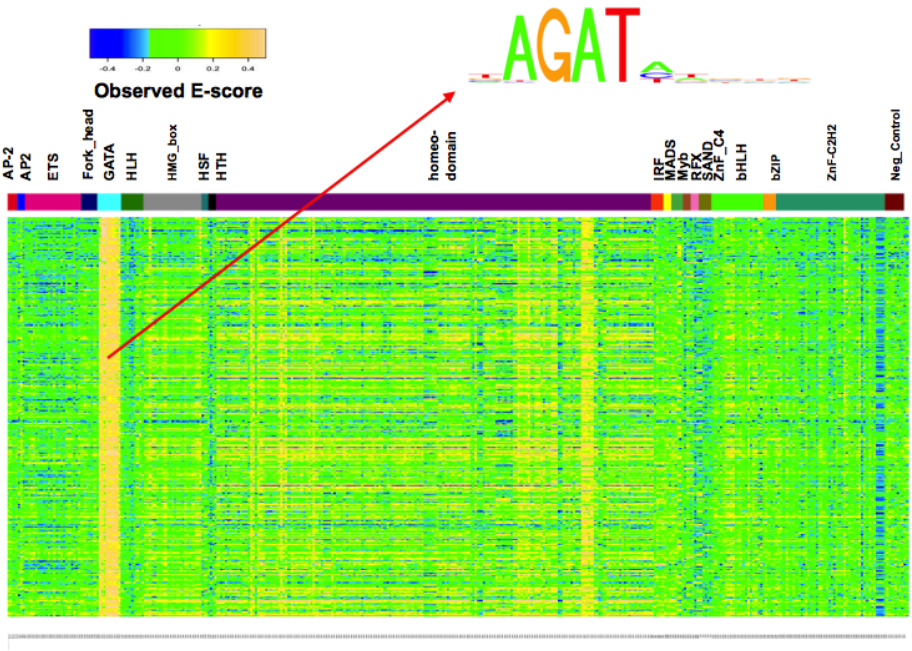


Figure 4.8: Characterization of TF-common $k$-mers for the GATA DBD class and their corresponding E-scores across our PBM data sets. The multi-colored strip above the heatmap is as in Figure 4.3

which appears to provide for indirect binding of the ChIP-profiled TF (*i.e.*, Yap6) to DNA (Gordân et al., 2009).

### 4.4.4 Identification of TF-common $k$-mers

By using the ANOVA decomposition model, we are able to identify groups of $k$-mers bound in common across the family ('TF- common' $k$-mers). For example, $k$-mers with high posterior probabilities for being TF-common have uniformly high E-scores for TFs in the ETS DBD class and relatively low E-scores for TFs in all the other DBD classes (Figure 4.12a). The PWM constructed from these $k$-mers indicates a shared binding specificity for the ETS DBD class (Figure 4.12a, upper panel). Characterization of TF-common $k$-mers for other DBD classes (GATA, HLH and HMG-box) are given in Figure 4.8, 4.9 and 4.10.

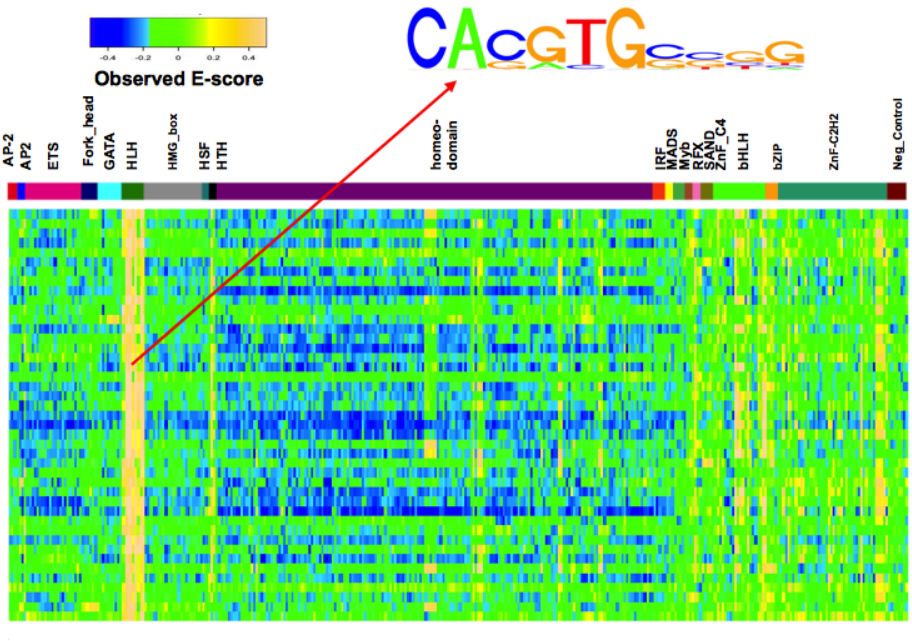


Figure 4.9: Characterization of TF-common $k$-mers for the HLH DBD class and their corresponding E-scores across our PBM data sets. The multi-colored strip above the heatmap is as in Figure 4.3

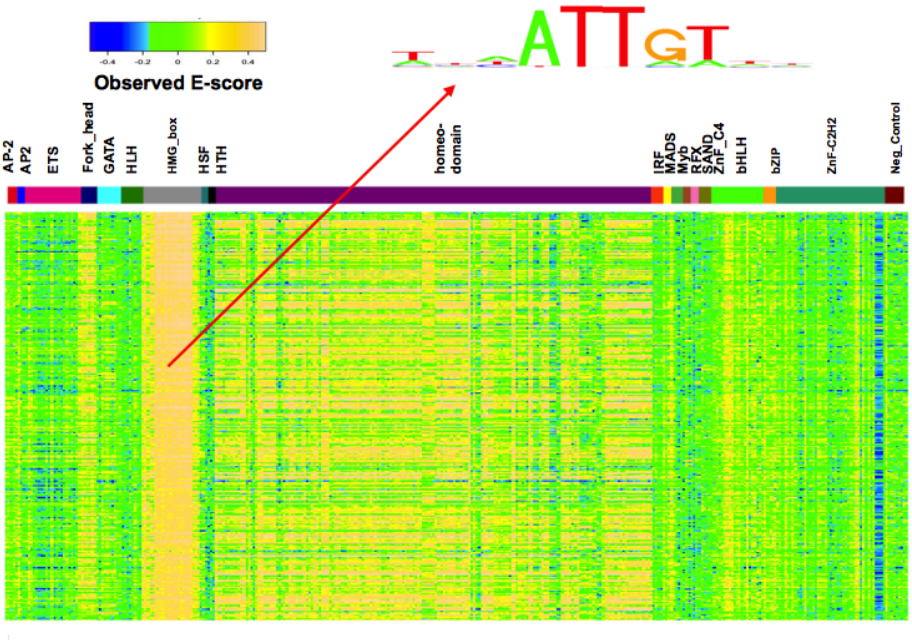


Figure 4.10: Characterization of TF-common $k$-mers for the HMG-box DBD class and their corresponding E-scores across our PBM data sets. The multi-colored strip above the heatmap is as in Figure 4.3

Posterior distribution of family-wise effects and numbers of TF-common $k$-mers for all the DBD classes are given in Figure 4.11.

### 4.4.5 Identification of TF subclasses based on similarity of PBM $k$-mer data

Our Bayesian hierarchical partition model allows for categorization of TF subclasses based on DNA binding preferences, which can simultaneously determine common binding sequences for each subclass. Previously, Berger et al. discovered separate DNA-binding specificity subgroups by considering the overlap among top 100 highest-affinity 8-mers for homeodomains (Berger et al., 2008). Distinct binding patterns were also identified by manually examining 8-mers with E-scores greater than a threshold score of 0.45. Our model-based analysis of homeodomains not only shows subclassification that is consistent with the results of Berger et al. (2008), but also systematically characterizes the common binding sequences for different subclasses of homeodomains (see Figure 4.13).

We further applied our model to determine subclasses and their DNA binding specificities in the ETS DBD class. Classification of 22 mouse ETS factors by hierarchical clustering (Figure 4.12c) over two groups of 8-mers identified in our analysis as being preferred by ETS subclasses is in general consistent with the classification obtained by aligning ETS-domain peptide sequences using the ClustalW algorithm (Figure 4.12d), and is broadly similar to the results obtained in (Wei et al., 2010), in which the similarity between DNA binding specificity motifs was obtained using the minimum Kullback-Leibler divergence between the multinomial distributions defined by the motifs. Notably, motifs generated according to subclass-preferred 8-mers (Figure 4.12c) are different from the motif generated from the TF-

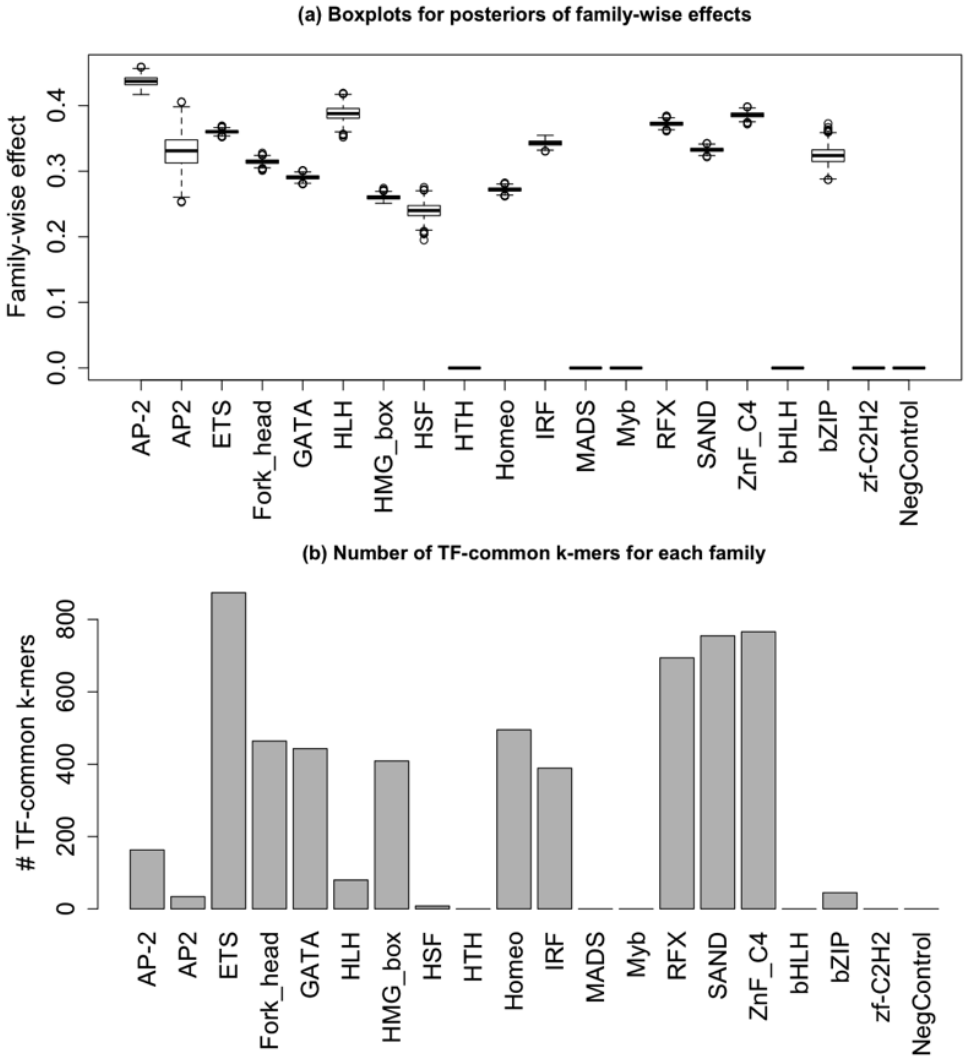


Figure 4.11: (a) Boxplots for posteriors of family-wise effects from different DBD classes. (b) Number of TF-common $k$-mers with posterior probability larger than 0.5 for different DBD classes. Five DBD classes (HTH, MADS, Myb, bHLH and Zf-C2H2) and the negative control group (NegControl) do not have any TF-common $k$-mer and their family-wise effects are zero.

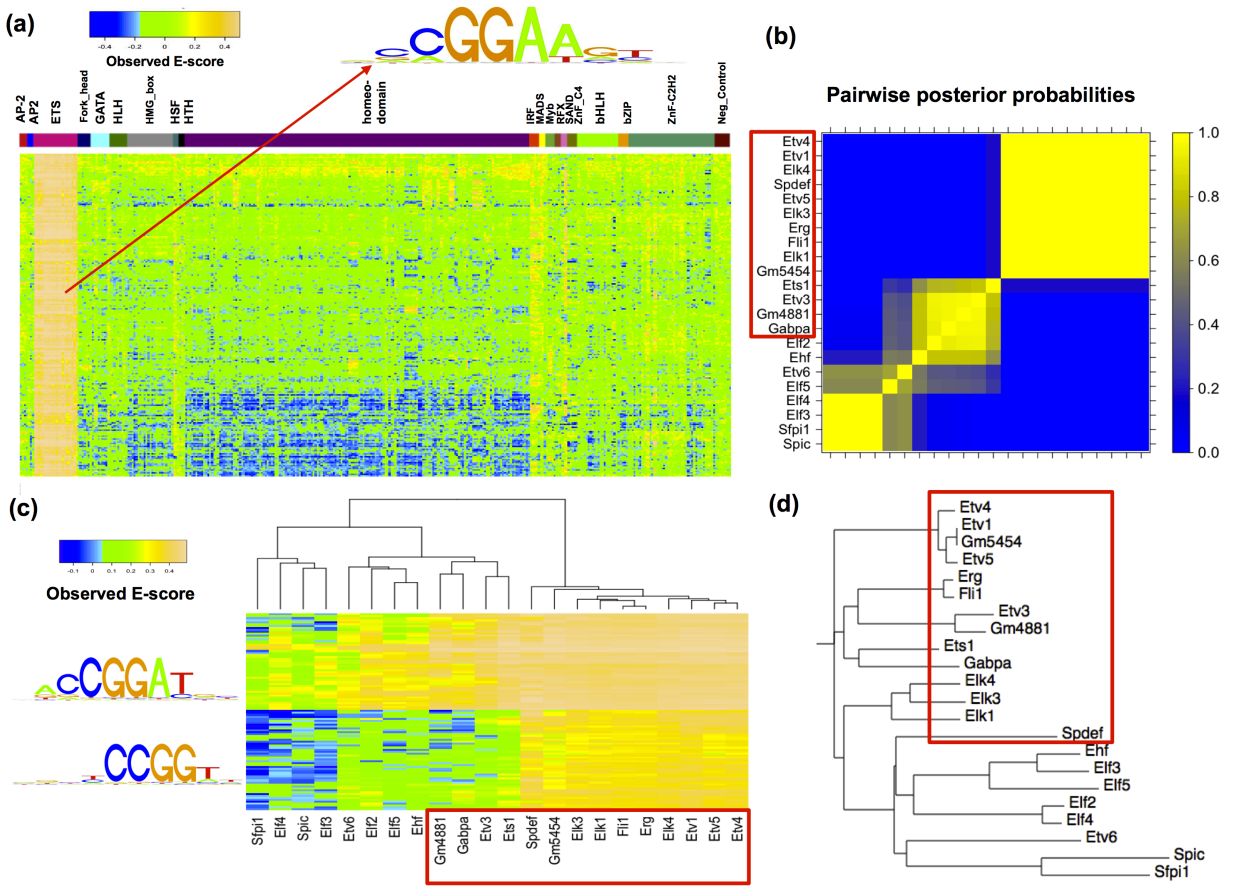


Figure 4.12: Categorization of TF subclasses for the ETS DBD class. (a) TF-common $k$-mers for the ETS DBD class and their corresponding E-scores across our PBM data sets. The multi-colored strip above the heatmap is as in Figure 4.3 (b) Pairwise posterior probabilities for clustering 22 mouse ETS factors based on the partition model described in Section 4.3.1. (c) Hierarchical clustering of 22 mouse ETS factors based on two groups of 8-mers identified as subclass-preferred. (d) Classification of members of the ETS DBD class by aligning ETS-domain peptide sequences using the ClustalW algorithm. TFs marked with red boxes in (b), (c) and (d) show a strong preference for the core sequence GGAT relative to other members of the ETS DBD class.

common $k$-mers (Figure 4.12a). A subclass of the ETS factors with similar ETS-domain peptide sequences according to ClustalW shows a specific binding preference to a consensus sequence ACCGGAT (marked by a red box in Figure 4.12c and 4.12d). Interestingly, members of this subclass can be distinguished further according to their binding preferences for the consensus sequence CCGGT. Differential binding preference by ETS factors for the core sequence GGAT has been observed previously (Wei et al., 2010), where its molecular basis was explored. By automatically identifying 8-mers that have the most distinct binding patterns, our model is able to characterize the binding specificities among members of the ETS DBD class in more detail.

### 4.4.6 Identification of TF-preferred $k$-mers

Highly similar members of a TF family can show different DNA sequence binding preferences. For example, the homeodomains Lhx4 and Lhx2 both bind most preferentially to the canonical sequence TAATTA, but differ in their preferences for other $k$-mers (Berger et al., 2008). In Figure 4.14a, we show the DNA binding specificity motifs for 8-mers that are identified as TF-preferred by Lhx2 but not Lhx4, and for those identified as TF-preferred by Lhx4 but not Lhx2, according to the ANOVA model described in Section 4.2.3. At the same time, we also applied the procedure described in Section 4.2.2 to search for TF-preferred $k$-mers based on their statistical significance calculated by an intersection-union test. In order to have a sufficiently large sample size, we focused on TF-preferred 6-mers in this study, and our tests are based on observed E-scores of 8-mers containing a given 6-mer. Analyses based on observed E-scores yielded nearly identical results as those based on corrected E-scores since the differences between E-scores on a pair experiments for the same $k$-mer are invariant to correction. The top three TF-preferred 6-mers (each at $P$-value $< 1.0 \times 10^{-7}$) found by

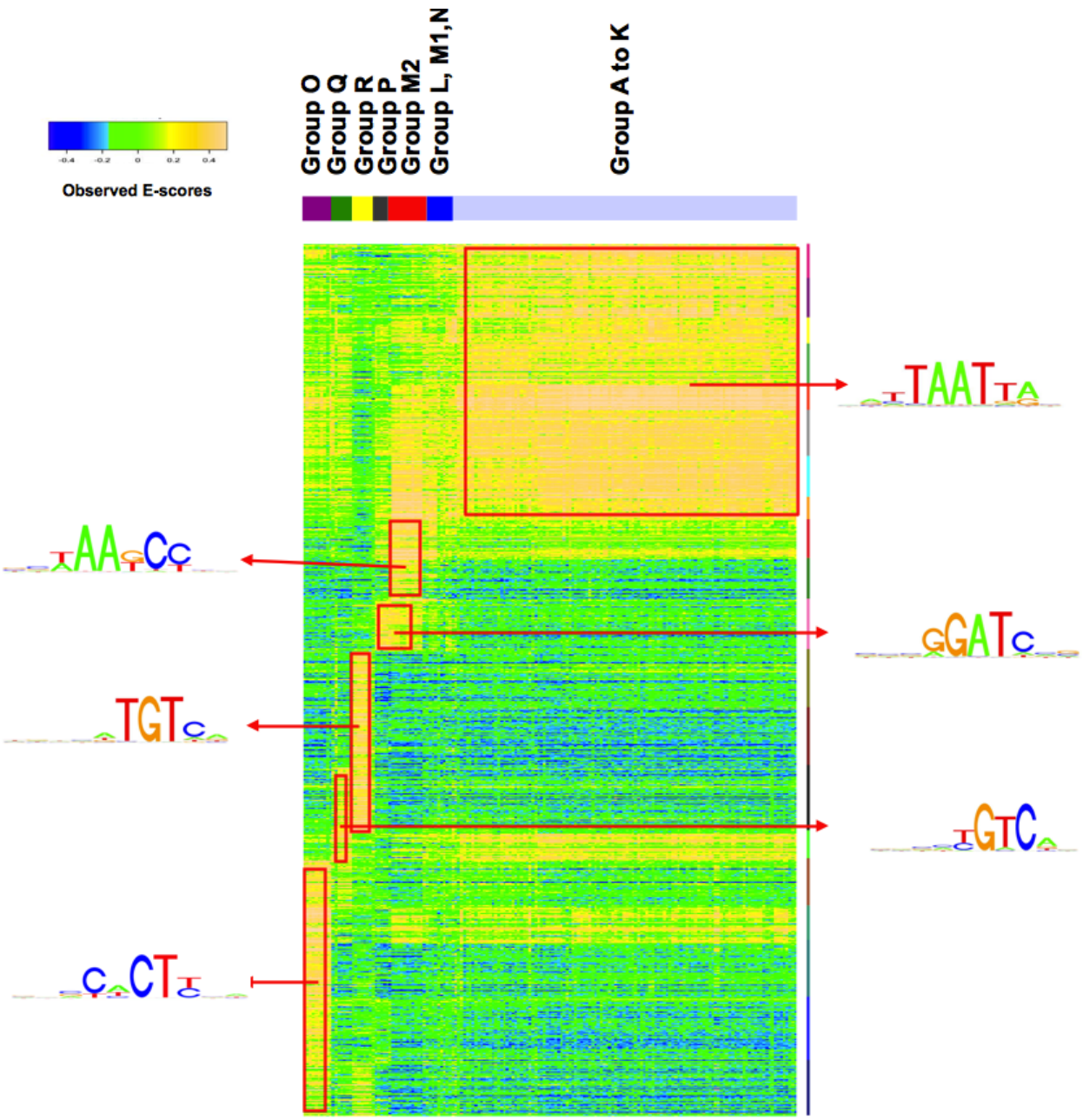


Figure 4.13: Subclassification of the homeodomain DBD class according to subclass-common 8-mers. 173 homeodomain TFs are classified into 7 subclasses (indicated by multi-colored strip on top of the heatmap), which are consistent with sub-groups identified in Berger et al. (2008) (denoted as Group A-R). Motif logos are generated according to the algorithm described in Section 4.3.2 by using subclass-common 8-mers (with equal weights).

this analysis in a pairwise comparison of Lhx2 and Lhx4 are apparent as off-diagonal points in a scatter-plot of 8-mer E-scores (Figure 4.14b). Of note, the TF-preferred 6-mers identified by our model-based approach are consistent with TF-preferred 6-mers identified by the intersection-union test. All four 6-mers (TAATGA and TAACGA for Lhx2, and TAATCA and TAATCT for Lhx4) identified in Berger, et al., 2008 by a primarily manual approach are also significant in our new, automated analysis (Figure 4.15). In addition, our automated analysis finds additional TF-preferred $k$-mers; for example, it finds TAATGG as a statistically significant TF-preferred 6-mer ($P$-value $= 6.9 \times 10^{-17}$) for Lhx2, and CAATCA as statistically significant 6-mer ($P$-value $= 1.2 \times 10^{-23}$) for Lhx4, in a pairwise comparison of those two TFs. Note that our analysis identifies CAATCA as preferred by Lhx4 in comparison with both Lhx2 and also with Lhx3 (Figure 4.16).

## 4.5 Discussion

Accurate, high-resolution datasets on the binding preferences of TFs for comprehensive collections of DNA sequences are essential for understanding the nature of protein-DNA binding specificity and how those specificities are used in transcriptional regulatory codes encoded in genomes. In this study, we developed a Bayesian model-based approach for analyzing $k$-mer TF DNA binding specificity data obtained from universal PBM experiments (Berger et al., 2006). Our model decomposes $k$-mer data (here, 8-mers) into artifactually high-scoring 8-mers, 8-mers bound in common by a TF family, and those bound preferentially by a particular member(s) of a TF family (TF-preferred $k$-mers). Adjusting PBM 8-mer E-scores for the identified systematic biases improved overall PBM data quality and correlations with *in vivo* TF binding data obtained by ChIP-chip (Harbison et al., 2004). The TF

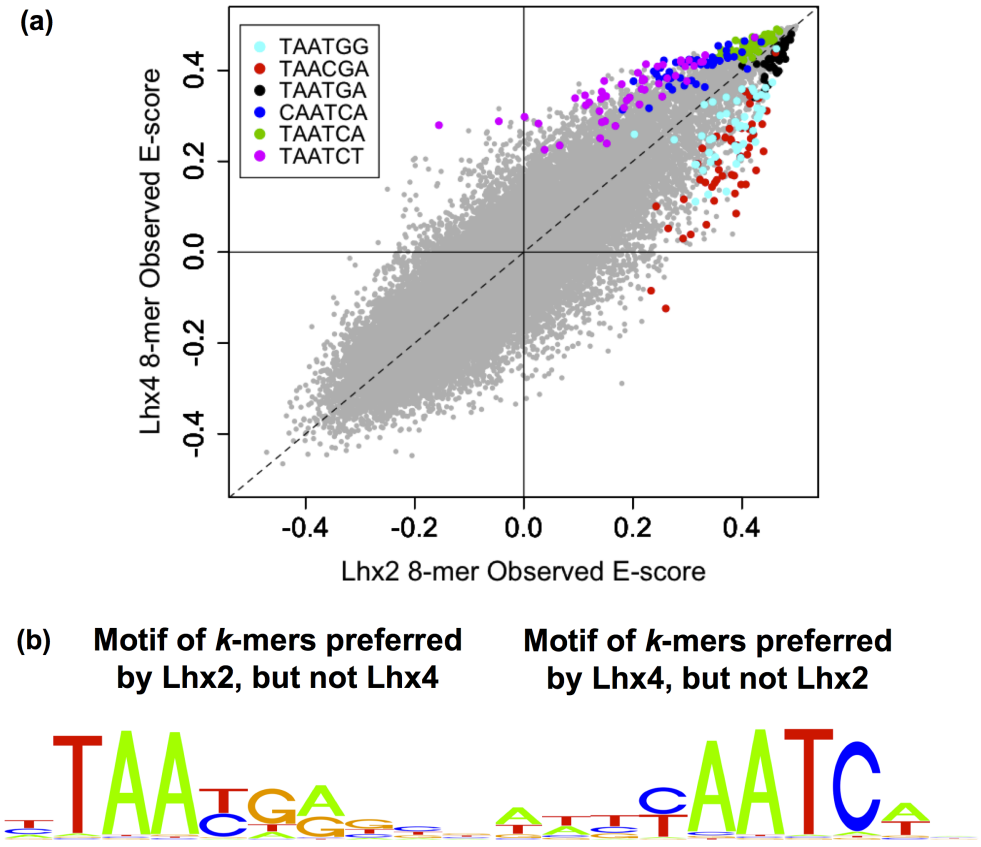


Figure 4.14: Model-based identification of TF-preferred $k$-mers. (a) Scatter plot of 8-mer E-scores comparing Lhx2 and Lhx4. 8-mers containing each of top three most significantly TF-preferred 6-mers from a direct comparison of Lhx2 and Lhx4 are highlighted in colors, revealing clear systematic differences in the binding by Lhx2 or Lhx4 to these sequences. (b) Sequence motif logos of TF-preferred 8-mers for Lhx2 (left) and Lhx4 (right). Sequence motifs were generated as described in Section 4.2.3 using 8-mers (with equal weights) that are identified by the ANOVA model as TF-preferred by one TF but not the other.

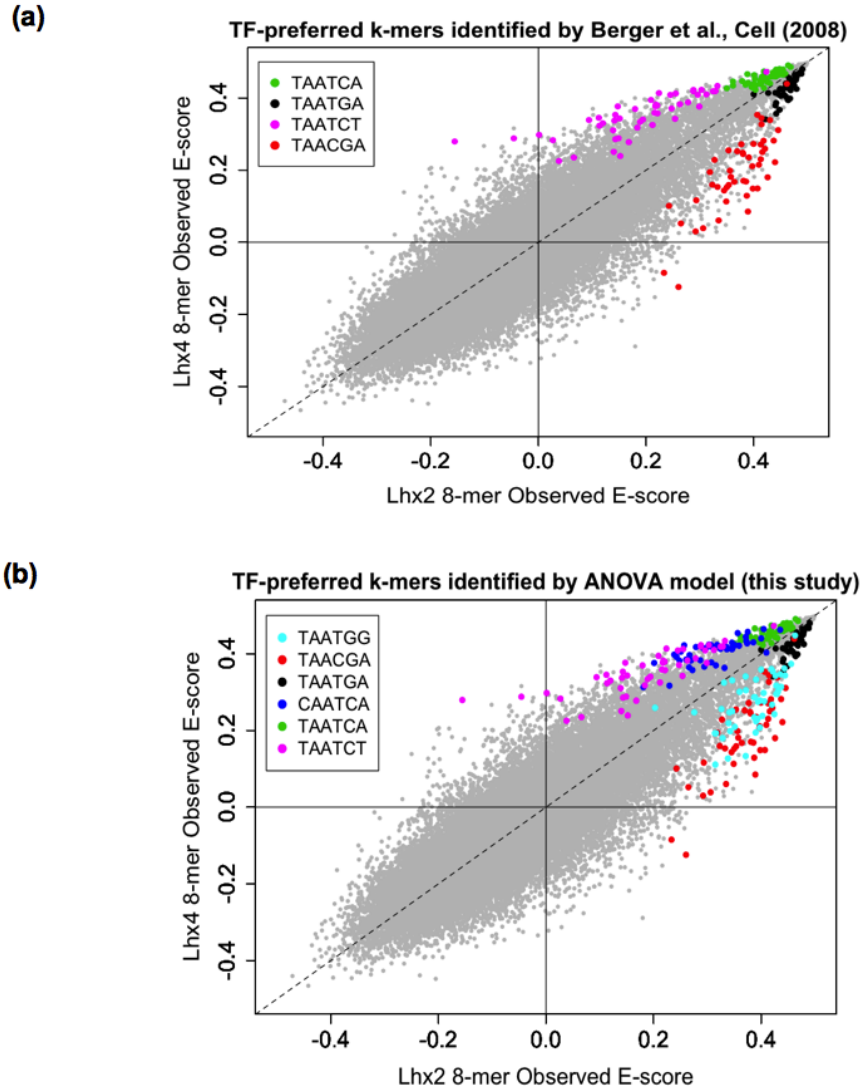


Figure 4.15: TF-preferred analysis for Lhx2 versus Lhx4. (a) Original results in Berger et al. (2008). (b) Results based on TF-preferred analysis from the ANOVA model described in this Jiang et al. paper. Note that we identified additional 6-mer 'CAATCA' to be preferred by Lhx4 and 6-mer 'TAATGG' to be preferred by Lhx2.

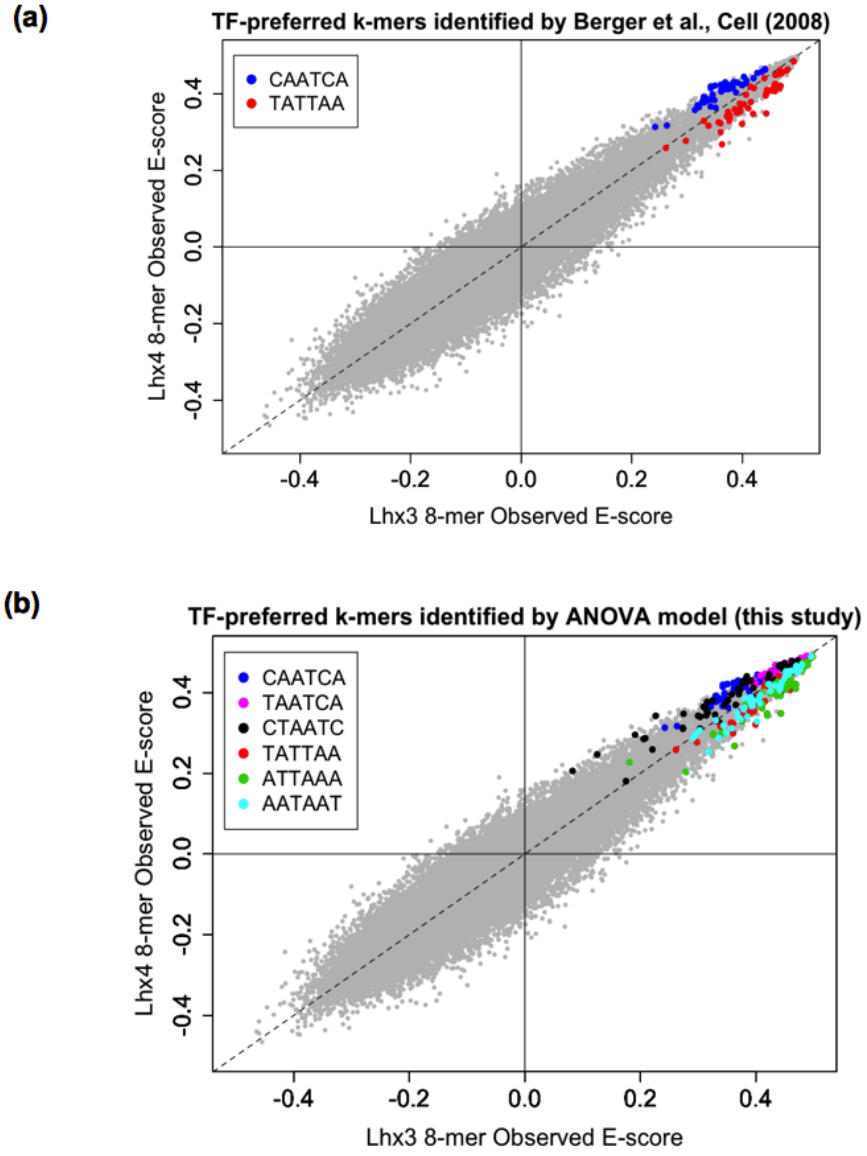


Figure 4.16: TF-preferred analysis for Lhx3 vs Lhx4. (a) Original results in Berger et al. (2008). (b) Results based on TF-preferred analysis from the ANOVA model described in this Jiang et al. paper. Note that we identified additional 6-mers 'TAATCA' and 'CTAATC' to be preferred preferred by Lhx4 and 6-mers 'ATTAAA' and 'AATAAT' to be preferred by Lhx3.

subclasses identified by our modeling approach are consistent with TF sub- classes based on TF DBD protein sequence similarity. TF-preferred $k$-mers are identified in an automated and systematic fashion by our model, without relying on visual inspection, manual curation, or arbitrary thresholds; our model captures TF-preferred $k$-mers previously identified through a combination of such other methods (Berger et al., 2008; Busser et al., 2012), while being more comprehensive in identifying statistically significantly TF-preferred $k$-mers. Systematic identification of TF-preferred $k$-mers should help to reduce investigator bias in searching for TF-preferred $k$-mers, and should aid in studies aimed at investigating the potential regulatory significance of TF-preferred versus TF- common $k$-mers (Gordân et al., 2011; Busser et al., 2012).

While our analysis identified 8-mers that tend to score artifactually highly in the universal PBM data sets that we examined, on its own it does not provide an explanation for these observations. We investigated the various data normalizations that are performed on the universal PBM data, but did not find any of them to contribute to artifactually high scores for these 8-mers. We cannot exclude the possibility that these 'sticky' 8-mers constitute a distinct set of nonspecific sequences that are bound by numerous TFs more preferentially than truly nonspecific or even disfavored sequences. Determining the underlying cause of these 'sticky' 8-mers will require additional experimental studies in the future.

In order to distinguish family-wise binding effects from systematic biases, our ANOVA model (4.1) requires a collection of TFs from diverse DBD classes and a sufficient number of TFs from each DBD class (at least three in this study). Estimation of $k$-mer background noise given a limited number of experiments (e.g., when adopting a new array design or platform) can be challenging. In this study, we focused our analysis on the observed E-scores due to the robustness of the E-score to experimental variation (Berger et al., 2006).

Future development of non-rank-based approaches might allow for improved classification of TFs and $k$-mers.

In this study, we analyzed data from two specific universal array designs, synthesized based on two different de Bruijn sequences, each of which covers all 10-mers. Our model could be applied to $k$-mer data generated using other universal array designs, including those based on higher-order de Bruijn sequences that comprehen- sively cover longer $k$-mers (Philippakis et al., 2008). Moreover, our approach is not limited to array designs based on de Bruijn sequences, but rather can be applied to any data sets using PBMs or other assays for which binding scores for $k$-mers are generated.

Numerous studies have focused on different TF structural classes, with the goal of identifying recognition rules underlying protein-DNA binding specificity (Suzuki and Yagi, 1994; Benos et al., 2002; Noyes et al., 2008; De Masi et al., 2011). Precise classification of TFs according to their DNA binding sequence preferences together with identification of those sets of preferred sequences, as provided by our modeling approach, will permit more detailed studies of the molecular determinants of TF-DNA binding specificity. Improved identification of $k$-mers bound preferentially by different TF family members will aid in investigations of what amino acid residues in the proteins correlate with differences in preferences for binding different $k$-mers.

Many studies of DNA regulatory elements have searched for combinations of motifs enriched within known or putative *cis* regulatory elements (Warner et al., 2008), including investigations of whether there are preferential spacings or orientations of how the TF binding sites are arranged within promoters (Beer and Tavazoie, 2004; Senger et al., 2004) or transcriptional enhancers (Fakhouri et al., 2010). Moreover, how different TF family members achieve their distinct regulatory effects is still not well understood; TF-preferred $k$-mers

constitute one mechanism by which paralogous TFs can attain distinct regulatory roles (Hollenhorst et al., 2009; Busser et al., 2012; Fong et al., 2012). More accurate, precise data on the DNA binding sequence preferences of different TFs, in particular paralogous TFs, will be important for more detailed investigations of *cis* regulatory codes.

The Bayesian modeling approach we present in this study is general and could be applied to other data types, beyond DNA-binding specificity data. Our modeling approach could be adapted to other sequence or experimental data sets in order to identify data features that are common to classes of proteins, defined according to either DBD structural class as we did in this study for sequence-specific TFs or to other annotations which may be more relevant for other types of proteins, versus features that are specific to individual proteins or subsets of proteins. Results from such studies might contribute to an improved understanding of different families of proteins, including the redundant versus divergent functions of individual members of protein families that arose from ancient gene duplications.

# Chapter 5

# Conclusion

This thesis has presented three new research contributions for variable selection methods and their applications.

First, we developed an effective and computationally efficient procedure for detecting interactions from an inverse modeling perspective. We found interesting links with sliced inverse regression (Li, 1991) and correlation pursuit (Zhong et al., 2012). The proposed method can also be viewed as a Frequentist analogue of the Bayesian partition model (Zhang et al., 2010) for continuous response and predictor variables. Frequentist properties of the proposed procedure are established and, in particular, its variable selection performance under a diverging number of predictors and sample size is investigated.

Second, we proposed a sequential partition prior and a dynamic slicing scheme for building a Bayesian partition model (Zhang et al., 2010) in eQTL studies. We augmented the partition to capture complex dependence structure among gene expression and linkage disequilibrium between genetic markers. Compared with the original model proposed by Zhang et al. (2010), the augmented model achieved significantly higher power in detecting genetic markers that

have epistatic interaction effects on multiple sets of co-regulated genes.

Third, we generalized the partition model to variable selection in the context of unsupervised learning. We were particularly motivated by its application in identifying transcription factor (TF) families based on DNA binding preferences determined by protein binding microarray (PBM) experiments. By introducing latent clusters (partitions), a Bayesian model was used to simultaneously identify TF families and short DNA sequences that are bound by most of the TFs in a family. We further developed a Bayesian analysis of variance (ANOVA) model that decomposes PBM binding scores into background noise, TF family-wise effects, and effects due to the particular TF. We showed that adjusting for background noise improved PBM data quality and concordance with *in vivo* TF binding data.

As data collection technology and data storage devices become more powerful, high-dimensional data are becoming increasingly common in many scientific fields. Variable selection methods play important roles in statistical modeling of high-dimensional data and are keys to data-driven discoveries in health, physical, and social sciences. The Naïve Bayes modeling approach and the partition model that combines forward and inverse perspectives are useful analysis tools for revealing and extracting relevant information from *big data*. While Frequentist methods such as SIRI usually enjoy theoretical guarantees and computational simplicity, the study of partition models under the Bayesian framework is particularly attractive because of its flexibility in dealing with latent (missing) variables and ability to incorporate prior knowledge.

The well known statistician George Box put it well when he said "all models are wrong but some models are useful". Although development of 'generic' variable selection methods provide convenient off-the-shelf tools for practitioners, continued research on partition models for variable selection can be most useful by solving real-world problems. Here are some

examples of possible directions for further research on partition models:

1. Bayesian partition model for continuous response and predictor variables. This can be viewed as a Bayesian analogue of the SIRI procedure. Other than computational costs, it requires problem-specific knowledge to choose background distributions of irrelevant predictors and their priors. For example, the success of the Bayesian partition model in eQTL studies partly relies on our knowledge of linkage disequilibrium and the block structure of correlations among genetic markers.

2. Independent screening procedure in eQTL studies. The inverse modeling perspective can be used to obtain simple screening statistics for filtering important genetic markers with either marginal or interaction effects. For example, under the multinomial model of genotypes, we can use the dynamic slicing scheme to find the maximized likelihood-ratio test statistic, which is equivalent to a maximized mutual information metric. We can also introduce priors for multinomial parameters and the *maximum a posteriori* (MAP) statistics may be less sensitive to markers with small minor allele frequencies. The theoretical properties and empirical performances of the proposed statistics are under study.

3. Bayesian partition model for studying treatment and covariate interactions under the potential outcome framework (Rubin, 1978, 2005). To develop strategies for personalized medicine, it is important to identify the treatment and covariate interactions in the setting of randomized clinical trial (Royston and Sauerbrei, 2008). Coupled with the missing data mechanism under the potential outcome framework, the Bayesian partition model can be useful for selecting important covariates associated with varying responses under different treatments.

# Appendix A

# Detailed Proofs

## A.1 Proof of Properties of Likelihood-Ratio Test Statistic in Section 2.1.1

Given the set of relevant predictors indexed by $\mathcal{A}$ with size $|\mathcal{A}|$ in model (2.3), let $\mathbf{x}_{\mathcal{A}}$ denote a $n \times |\mathcal{A}|$ matrix of observed variables in index set $\mathcal{A}$, and $\mathbf{x}_{\mathcal{A}}^{hj}$ $(1 \leq j \leq n_h)$ denote a $|\mathcal{A}|$-dimensional column vector of variables in index set $\mathcal{A}$ for $j$th observation in slice $S_h$ $(1 \leq h \leq H)$. We denote $B_{\mathcal{A}} = \mathrm{Cov}\left(\mathbb{E}(\mathbf{X}_{\mathcal{A}}|S(Y))\right)$, $W_{\mathcal{A}} = \mathbb{E}\left(\mathrm{Cov}(\mathbf{X}_{\mathcal{A}}|S(Y))\right)$ and $\Omega_{\mathcal{A}} = B_{\mathcal{A}} + W_{\mathcal{A}}$. The corresponding sample estimates are given by

$$\widehat{B}_{\mathcal{A}} = \frac{1}{n}\sum_{h=1}^{H} n_h \left(\bar{\mathbf{x}}_{\mathcal{A}}^{h\cdot} - \bar{\mathbf{x}}_{\mathcal{A}}^{\cdot\cdot}\right)\left(\bar{\mathbf{x}}_{\mathcal{A}}^{h\cdot} - \bar{\mathbf{x}}_{\mathcal{A}}^{\cdot\cdot}\right)^T, \tag{A.1}$$

$$\widehat{W}_{\mathcal{A}} = \frac{1}{n}\sum_{h=1}^{H}\sum_{j=1}^{n_h}\left(\mathbf{x}_{\mathcal{A}}^{hj} - \bar{\mathbf{x}}_{\mathcal{A}}^{h\cdot}\right)\left(\mathbf{x}_{\mathcal{A}}^{hj} - \bar{\mathbf{x}}_{\mathcal{A}}^{h\cdot}\right)^T,$$

and $\widehat{\Omega}_{\mathcal{A}} = \widehat{B}_{\mathcal{A}} + \widehat{W}_{\mathcal{A}}$, where $\bar{\mathbf{x}}_{\mathcal{A}}^{\cdot\cdot} = \frac{1}{n}\sum_{h=1}^{H}\sum_{j=1}^{n_h}\mathbf{x}_{\mathcal{A}}^{hj}$ and $\bar{\mathbf{x}}_{\mathcal{A}}^{h\cdot} = \frac{1}{n_h}\sum_{j=1}^{n_h}\mathbf{x}_{\mathcal{A}}^{hj}$.

Szretter and Yohai (2009) proved the following proposition:

**Proposition 3.** *Let $C$ be the orthogonal matrix $[\mathbf{c}_1, \ldots, \mathbf{c}_p]$, where $\mathbf{c}_j$ is an eigenvector of $\widehat{B}_{\mathcal{A}}^{-1/2}\widehat{W}_{\mathcal{A}}\widehat{B}_{\mathcal{A}}^{-1/2} = C\Theta C^T$ corresponding to the eigenvalue $\theta_j$, where $\theta_1 \geq \theta_2 \geq \ldots \geq \theta_p$. $\Theta$ is the diagonal matrix with $\theta_1, \theta_2, \ldots, \theta_p$ in the diagonal. $C_r$ is the matrix with the first $r$ columns of $C$.*

*(a) The maximum likelihood estimate of $\Sigma = \Sigma_{\mathcal{A}} = \mathrm{Cov}\left(\mathbf{X}_{\mathcal{A}}|S(Y)\right)$ in model (2.3) is*

$$\widehat{\Sigma}_{\mathcal{A}} = \widehat{W}_{\mathcal{A}} + \widehat{B}_{\mathcal{A}}^{1/2} C_{p-q} C_{p-q}^T \widehat{B}_{\mathcal{A}}^{1/2}. \tag{A.2}$$

*(b) Let $\mathbf{u}_i$, $1 \leq i \leq p$, be orthogonal eigenvectors of norm one of $\widehat{\Sigma}_{\mathcal{A}}^{-1/2}\widehat{B}_{\mathcal{A}}\widehat{\Sigma}_{\mathcal{A}}^{-1/2}$ corresponding to the eigenvalues $\omega_1 \geq \omega_2 \geq \ldots \geq \omega_p$. The maximum likelihood estimate of $\mu_h = \mu_{\mathcal{A}}^{(h)} = \mathbb{E}\left(\mathbf{X}_{\mathcal{A}}|Y \in S_h\right)$ in model (2.3) is*

$$\widehat{\mu}_{\mathcal{A}}^{(h)} = \widehat{\Sigma}_{\mathcal{A}}^{1/2} U_K U_q^T \widehat{\Sigma}_{\mathcal{A}}^{-1/2} \left(\bar{\mathbf{x}}_{\mathcal{A}}^{h\cdot} - \bar{\mathbf{x}}_{\mathcal{A}}^{\cdot\cdot}\right) + \bar{\mathbf{x}}_{\mathcal{A}}^{\cdot\cdot}, 1 \leq h \leq H,$$

*where $U_q = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_q]$. Then, $\widehat{\mu}_{\mathcal{A}}^{(h)}$ is the orthogonal projection, using the norm associated to $\widehat{\Sigma}_{\mathcal{A}}$, of $\left(\bar{\mathbf{x}}_{\mathcal{A}}^{h\cdot} - \bar{\mathbf{x}}_{\mathcal{A}}^{\cdot\cdot}\right)$ on the $q$-dimensional affine subspace $\bar{\mathbf{x}}_{\mathcal{A}}^{\cdot\cdot} + \mathbb{V}^q$, where $\mathbb{V}^q$ is spanned by $\left(\widehat{\Sigma}_{\mathcal{A}}^{1/2}\mathbf{u}_1, \widehat{\Sigma}_{\mathcal{A}}^{1/2}\mathbf{u}_2, \ldots, \widehat{\Sigma}_{\mathcal{A}}^{1/2}\mathbf{u}_q\right)$.*

*(c) $\widehat{\Sigma}_{\mathcal{A}}$ can also be written as*

$$\widehat{\Sigma}_{\mathcal{A}} = \frac{1}{n}\sum_{h=1}^{H}\sum_{j=1}^{n_h}\left(\mathbf{x}_{\mathcal{A}}^{hj} - \widehat{\mu}_{\mathcal{A}}^{(h)}\right)\left(\mathbf{x}_{\mathcal{A}}^{hj} - \widehat{\mu}_{\mathcal{A}}^{(h)}\right)^T.$$

*$\widehat{\Sigma}_{\mathcal{A}}^{1/2}\mathbf{u}_j$ is an eigenvector of $\widehat{B}_{\mathcal{A}}\widehat{W}_{\mathcal{A}}^{-1}$ corresponding to eigenvalue $1/\theta_{p-i+1}$, $1 \leq i \leq p$.*

Properties of the likelihood-ratio test statistic proposed in Section 2.1.1 are summarized in the following proposition:

**Proposition 4.** *Given the current set of selected predictors indexed by $\mathcal{C}$ with dimension $|\mathcal{C}| = d$ and another predictor indexed by $j \notin \mathcal{C}$, the scaled log-likelihood-ratio test statistic for testing*

$$H_0 : \mathcal{A} = \mathcal{C} \text{ v.s. } H_1 : \mathcal{A} = \mathcal{C} \cup \{j\},$$

*can be written as*

$$\widehat{D}_{j|\mathcal{C}} = \frac{2}{n} \log\left(L_{j|\mathcal{C}}\right) = \sum_{k=1}^{q} \log\left(1 + \frac{\widehat{\lambda}_k^{d+1} - \widehat{\lambda}_k^{d}}{1 - \widehat{\lambda}_k^{d+1}}\right), \tag{A.3}$$

*where $q$ is the dimension of subspace $\mathbb{V}^q$ defined in model (2.3), $\lambda_k^d$ and $\lambda_k^{d+1}$ are the $k$th largest eigenvalues of $\Omega_{\mathcal{C}}^{-1} B_{\mathcal{C}}$ and $\Omega_{[\mathcal{C}\cup\{j\}]}^{-1} B_{[\mathcal{C}\cup\{j\}]}$, respectively, and $\widehat{\lambda}_k^d$ and $\widehat{\lambda}_k^{d+1}$ are their estimates. For any fixed slicing scheme and the true relevant predictors indexed by $\mathcal{A}$, let $\lambda_k$ be the $k$th largest eigenvalue of $\Omega_{\mathcal{A}}^{-1} B_{\mathcal{A}}$. We further assume that $\lambda_1 > \lambda_2 > \ldots > \lambda_q > 0$.*

*(a) Under the assumption that $\mathcal{A} \subset \mathcal{C}$,*

$$n\widehat{D}_{j|\mathcal{C}} \simeq \sum_{i=1}^{q} \frac{n\left(\widehat{\lambda}_i^{d+1} - \widehat{\lambda}_i^{d}\right)}{1 - \widehat{\lambda}_i^{d+1}}, \tag{A.4}$$

*which is asymptotically equivalent to the correlation pursuit (COP) test statistic defined in Zhong et al. (2012) and has an asymptotic distribution of $\chi_q^2$.*

*(b) Under the same conditions as in (a),*

$$\left(n\widehat{D}_{j|\mathcal{C}}\right)_{j\in\mathcal{C}^c} \xrightarrow{D} \left(\sum_{k=1}^{K} z_{kj}^2\right)_{j\in\mathcal{C}^c}, \text{ and } \max_{j\in\mathcal{C}^c}\left(n\widehat{D}_{j|\mathcal{C}}\right) \xrightarrow{D} \max_{j\in\mathcal{C}^c}\left(\sum_{i=1}^{q} z_{ij}^2\right),$$

*where $\mathbf{z}_k = (z_{kj})_{j\in\mathcal{C}^c} \sim MVN\left(\mathbf{0}, \left[Corr\left(X_i, X_j|\mathbf{X}_{\mathcal{C}}\right)\right]_{i,j\in\mathcal{C}^c}\right)$ and $\mathbf{z}_k$'s are independent.*

*(c) As* $n \to \infty$,

$$\widehat{D}_{j|\mathcal{C}} \xrightarrow{a.s.} D_{j|\mathcal{C}} = \log\left(1 + \frac{Var(M_j) - Cov\left(M_j, \mathbf{X}_{\mathcal{C}}\right)\left[Cov\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1} Cov\left(M_j, \mathbf{X}_{\mathcal{C}}\right)^T}{V_j}\right)$$

*where* $M_j = \mathbb{E}\left(X_j|\mathbf{X}_{\mathcal{C}}, S(Y)\right)$, $V_j = Var\left(X_j|\mathbf{X}_{\mathcal{C}}, S(Y)\right)$ *and* $S(Y) = h$ *when* $Y \in S_h$ *(*$1 \leq h \leq H$*).* $V_j$ *does not depend on* $Y$ *under the assumption in model (2.3). Furthermore,*

$$D_{j|\mathcal{C}} = 0 \text{ } \textit{iff} \, \mathbb{E}\left(X_j|\mathbf{X}_{\mathcal{C}}, Y \in S_h\right) = \mathbb{E}\left(X_j|\mathbf{X}_{\mathcal{C}}\right), 1 \leq h \leq H.$$

*Proof of Proposition 4.* One can show that

$$\widehat{D}_{j|\mathcal{C}} = \frac{2}{n}\log\left(L_{j|\mathcal{C}}\right) = -\left(\log\left[\det\left(\widehat{\Omega}^{-1}_{[\mathcal{C}\cup\{j\}]}\widehat{\Sigma}_{[\mathcal{C}\cup\{j\}]}\right)\right] - \log\left[\det\left(\widehat{\Omega}^{-1}_{\mathcal{C}}\widehat{\Sigma}_{\mathcal{C}}\right)\right]\right).$$

To prove (A.3), we just need to show that

$$\log\left[\det\left(\widehat{\Omega}^{-1}_{\mathcal{C}}\widehat{\Sigma}_{\mathcal{C}}\right)\right] = \sum_{i=1}^{q}\log\left(1 - \widehat{\lambda}^d_i\right),$$

where $d = |\mathcal{C}|$ and $\widehat{\lambda}^d_i$ is the $i$th largest eigenvalue of $\widehat{\Omega}^{-1}_{\mathcal{C}}\widehat{B}_{\mathcal{C}}$ corresponding to eigenvector $\boldsymbol{\eta}_i$, $1 \leq i \leq d$. Since $\widehat{B}_{\mathcal{C}}\boldsymbol{\eta}_i = \widehat{\lambda}^d_i\widehat{\Omega}_{\mathcal{C}}\boldsymbol{\eta}_i$ and $\widehat{\Omega}_{\mathcal{C}} = \widehat{W}_{\mathcal{C}} + \widehat{B}_{\mathcal{C}}$,

$$\widehat{\Omega}_{\mathcal{C}}\boldsymbol{\eta}_i = \frac{1}{\widehat{\lambda}^d_i}\widehat{B}_{\mathcal{C}}\boldsymbol{\eta}_i = \frac{1}{1 - \widehat{\lambda}^d_i}\widehat{W}_{\mathcal{C}}\boldsymbol{\eta}_i.$$

Then,

$$\widehat{W}_{\mathcal{C}}\widehat{B}^{-1}_{\mathcal{C}}\widehat{\Omega}_{\mathcal{C}}\boldsymbol{\eta}_i = \frac{1 - \widehat{\lambda}^d_i}{\widehat{\lambda}^d_i}\widehat{\Omega}_{\mathcal{C}}\boldsymbol{\eta}_i,$$

and

$$\left(\widehat{B}_{\mathcal{C}}^{-1/2}\widehat{W}_{\mathcal{C}}\widehat{B}_{\mathcal{C}}^{-1/2}\right)\widehat{B}_{\mathcal{C}}^{-1/2}\widehat{\Omega}_{\mathcal{C}}\boldsymbol{\eta}_i = \frac{1-\widehat{\lambda}_i^d}{\widehat{\lambda}_i^d}\widehat{B}_{\mathcal{C}}^{-1/2}\widehat{\Omega}_{\mathcal{C}}\boldsymbol{\eta}_i.$$

Thus, the eigenvalues of $\widehat{B}_{\mathcal{C}}^{-1/2}\widehat{W}_{\mathcal{C}}\widehat{B}_{\mathcal{C}}^{-1/2}$ are given by $\theta_{d-i+1} = \frac{1-\widehat{\lambda}_i^d}{\widehat{\lambda}_i^d}$, $1 \le i \le d$. Let $\mathbf{c}_i$ be an eigenvector of $\widehat{B}_{\mathcal{C}}^{-1/2}\widehat{W}_{\mathcal{C}}\widehat{B}_{\mathcal{C}}^{-1/2} = C\Theta C^T$ corresponding to the eigenvalue $\theta_i$. We will prove that the eigenvalues of $\widehat{\Sigma}_{\mathcal{C}}^{-1}\widehat{B}_{\mathcal{C}}$ are

$$\omega_i = \begin{cases} \frac{1}{\theta_{d-i+1}} = \frac{\widehat{\lambda}_i^d}{1-\widehat{\lambda}_i^d} & \text{if } 1 \le i \le q, \\ \frac{1}{1+\theta_{d-i+1}} = \widehat{\lambda}_i^d & \text{if } q+1 \le i \le d, \end{cases}$$

with corresponding eigenvectors given by $\mathbf{b}_i = \widehat{B}_{\mathcal{C}}^{-1/2}\mathbf{c}_{d-i+1}$. According to (A.2) in Proposition 3,

$$\begin{aligned} \widehat{B}_{\mathcal{C}}^{-1/2}\widehat{\Sigma}_{\mathcal{C}}\widehat{B}_{\mathcal{C}}^{-1/2}\mathbf{c}_{d-i+1} &= \widehat{B}_{\mathcal{C}}^{-1/2}\left(\widehat{W}_{\mathcal{C}} + \widehat{B}_{\mathcal{C}}^{1/2}C_{|\mathcal{A}|-q}C_{d-q}^T B_{\mathcal{C}}^{1/2}\right)\widehat{B}_{\mathcal{C}}^{-1/2}\mathbf{c}_{d-i+1} \\ &= \left(\widehat{B}_{\mathcal{C}}^{-1/2}\widehat{W}_{\mathcal{C}}\widehat{B}_{\mathcal{C}}^{-1/2} + C_{d-k}C_{d-q}^T\right)\mathbf{c}_{d-i+1}, \end{aligned}$$

where $\widehat{B}_{\mathcal{C}}^{-1/2}\widehat{W}_{\mathcal{C}}\widehat{B}_{\mathcal{C}}^{-1/2}\mathbf{c}_{d-i+1} = \theta_{d-i+1}$, $C_{d-q}C_{d-q}^T\mathbf{c}_{d-i+1}$ is 0 for $1 \le i \le q$ and 1 for $q+1 \le i \le d$. Thus,

$$\widehat{B}_{\mathcal{C}}^{-1/2}\widehat{\Sigma}_{\mathcal{C}}\widehat{B}_{\mathcal{C}}^{-1/2}\mathbf{c}_{d-i+1} = \frac{1}{\omega_i}\mathbf{c}_{d-i+1},$$

and

$$\widehat{\Sigma}_{\mathcal{C}}^{-1}\widehat{B}_{\mathcal{C}}\mathbf{b}_i = \omega_i\mathbf{b}_i, 1 \le i \le d.$$

Therefore,

$$\log\left[\det\left(\widehat{\Omega}_{\mathcal{C}}^{-1}\widehat{\Sigma}_{\mathcal{C}}\right)\right] = \log\left[\det\left(\widehat{\Omega}_{\mathcal{C}}^{-1}\widehat{B}_{\mathcal{C}}\right)\right] - \log\left[\det\left(\widehat{\Sigma}_{\mathcal{C}}^{-1}\widehat{B}_{\mathcal{C}}\right)\right] = \sum_{i=1}^{q}\log\left(1-\widehat{\lambda}_i^d\right).$$

To prove (a) and (b), note that

$$n\widehat{D}_{j|\mathcal{C}} = -n\left(\sum_{i=1}^{q}\log\left(1-\widehat{\lambda}_i^{d+1}\right)-\sum_{i=1}^{q}\log\left(1-\widehat{\lambda}_i^{d}\right)\right) = n\sum_{i=1}^{q}\log\left(1+\frac{\widehat{\lambda}_i^{d+1}-\widehat{\lambda}_i^{d}}{1-\widehat{\lambda}_i^{d+1}}\right).$$

Given that $\mathcal{A}\subset\mathcal{C}$, Zhong et al. (2012) showed that

$$\frac{n\left(\widehat{\lambda}_i^{d+1}-\widehat{\lambda}_i^{d}\right)}{1-\widehat{\lambda}_i^{d+1}}, i=1,2,\ldots,q,$$

are asymptotically independent and identically distributed as $\chi_1^2$. Thus,

$$\frac{\widehat{\lambda}_i^{d+1}-\widehat{\lambda}_i^{d}}{1-\widehat{\lambda}_i^{d+1}} \xrightarrow{P} 0, i=1,2,\ldots,q,$$

and

$$n\widehat{D}_{j|\mathcal{C}} = n\sum_{i=1}^{q}\log\left(1+\frac{\widehat{\lambda}_i^{d+1}-\widehat{\lambda}_i^{d}}{1-\widehat{\lambda}_i^{d+1}}\right) \simeq \sum_{i=1}^{q}\frac{n\left(\widehat{\lambda}_i^{d+1}-\widehat{\lambda}_i^{d}\right)}{1-\widehat{\lambda}_i^{d+1}}.$$

Conditioning on variables in $\mathcal{C}$, variables in $\mathcal{C}^c$ follows a multivariate normal distribution with the same mean and variance across slices. Then, the proofs of (a) and (b) directly follow from the Theorem 1 and Theorem 2 in Zhong et al. (2012).

For (c), since $\widehat{\lambda}_i^d \xrightarrow{P} \lambda_i^d$, for $i = 1, 2, \ldots, d$,

$$\begin{aligned}
\widehat{D}_{j|\mathcal{C}} &\xrightarrow{P} -\sum_{i=1}^{q} \log\left(1-\lambda_i^{d+1}\right) + \sum_{i=1}^{q} \log\left(1-\lambda_i^{d}\right) \\
&= -\log\left[\det\left(\Omega_{[\mathcal{C}\cup\{j\}]}^{-1} W_{[\mathcal{C}\cup\{j\}]}\right)\right] + \log\left[\det\left(\Omega_{\mathcal{C}}^{-1} W_{\mathcal{C}}\right)\right] \\
&= \log\left[\frac{\det\left(\Omega_{[\mathcal{C}\cup\{j\}]}\right)}{\det\left(\Omega_{\mathcal{C}}\right)}\right] - \log\left[\frac{\det\left(W_{[\mathcal{C}\cup\{j\}]}\right)}{\det\left(W_{\mathcal{C}}\right)}\right] \\
&= \log\left[\mathrm{Var}\left(X_j\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)^T\right] - \log\left(V_j\right) \\
&= \log\left(1 + \frac{\mathrm{Var}\left(M_j\right) - \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)^T}{V_j}\right).
\end{aligned}$$

The last equality follows from $\mathrm{Var}\left(X_j\right) = \mathrm{Var}\left(M_j\right) + \mathbb{E}(V_j) = \mathrm{Var}\left(M_j\right) + V_j$ and $\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right) = \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)$, where $M_j = \mathbb{E}\left(X_j | \mathbf{X}_{\mathcal{C}}, S(Y)\right)$, $V_j = \mathrm{Var}\left(X_j | \mathbf{X}_{\mathcal{C}}, S(Y)\right)$ and $V_j$ is a constant that does not depend on $\mathbf{X}_{\mathcal{C}}$ or $S(Y)$. Note that

$$\mathrm{Var}\left(M_j\right) \geq \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)^T$$

where equality holds if and only if $M_j = \mathbb{E}\left(X_j | \mathbf{X}_{\mathcal{C}}, S(Y)\right)$ is a linear combination of $\mathbf{X}_{\mathcal{C}}$ that does not depend on $S(Y)$, that is, $\mathbb{E}\left(X_j | \mathbf{X}_{\mathcal{C}}, Y \in S_h\right) = \mathbb{E}\left(X_j | \mathbf{X}_{\mathcal{C}}\right)$ for $1 \leq h \leq H$ under the normality assumption. □

## A.2 Proof of Theorem 1 in Section 2.1.1

To prove Theorem 1, we will need the following two lemmas.

**Lemma 1.** *Under the same conditions as in Theorem 1, there exist* $\kappa_1 > 0$ *and* $\xi_1 > 0$ *such*

*that for any set of predictors indexed by* $\mathcal{C}$ *and* $\mathcal{C}^c \cap \mathcal{A} \neq \emptyset$,

$$\max_{j \in \mathcal{C}^c \cap \mathcal{A}} \left[ \frac{Var(M_j) - Cov(M_j, \mathbf{X}_\mathcal{C}) \left[Cov(\mathbf{X}_\mathcal{C})\right]^{-1} Cov(M_j, \mathbf{X}_\mathcal{C})^T}{V_j} \right] \geq \xi_1 n^{-\kappa},$$

*where* $M_j = \mathbb{E}(X_j | \mathbf{X}_\mathcal{C}, S(Y))$, $V_j = Var(X_j | \mathbf{X}_\mathcal{C}, S(Y))$ *and* $V_j$ *is a constant that does not depend on* $\mathbf{X}_\mathcal{C}$ *or* $S(Y)$.

**Lemma 2.** *Under the same conditions as in Theorem 1, for any set of predictors indexed by* $\mathcal{C}$, *let* $\widehat{\lambda}_i^\mathcal{C}$ *be the* $i$*th largest eigenvalue of* $\widehat{\Omega}_\mathcal{C}^{-1} \widehat{B}_\mathcal{C}$ *and let* $\lambda_i^\mathcal{C}$ *be the* $i$*th largest eigenvalue of* $\Omega_\mathcal{C}^{-1} B_\mathcal{C}$. *Then, for* $0 < \epsilon < 1$ *and* $i = 1, 2, \ldots, q$, *there exists positive constants* $C_1$ *and* $C_2$ *such that*

$$Pr\left( \max_{\mathcal{C} \subset \{1,2,\ldots,p\}} \left| \log\left(1 - \widehat{\lambda}_i^\mathcal{C}\right) - \log\left(1 - \lambda_i^\mathcal{C}\right) \right| > \epsilon \right) \leq 2p(p+1) C_1 \exp\left( -C_2 n \frac{\tau_{min}^4 \epsilon^2}{64 \tau_{max}^2 p^2} \right)$$

We will start with the proofs of two lemmas.

*Proof of Lemma 1.* Let $\mathcal{B} = \mathcal{C} \cap \mathcal{A}$, $\mathcal{E} = \mathcal{C} \cap \mathcal{A}^c$, and $\mathcal{D} = \mathcal{C}^c \cap \mathcal{A} \neq \emptyset$. Under model (2.3),

$$\mathbf{X}_\mathcal{D} | \mathbf{X}_\mathcal{B}, Y \in S_h \sim N\left( \alpha_{\mathcal{D}|\mathcal{B}}^{(h)} + \boldsymbol{\beta}_{\mathcal{D}|\mathcal{B}}^T \mathbf{X}_\mathcal{B}, \Sigma_1 = \Sigma_{\mathcal{D}|\mathcal{B}} \right).$$

Let $\alpha_\mathcal{D}^{(h)} = \left( \alpha_{j \in \mathcal{D}}^{(h)} \right)^T$, where $\alpha_j^{(h)}$ is defined in (2.6). Then, $\alpha_\mathcal{D}^{(h)} = \Psi \alpha_{\mathcal{D}|\mathcal{B}}^{(h)}$, where $\Psi$ is a $|\mathcal{D}|$ by $|\mathcal{D}|$ matrix that satisfies $\Psi^T \Psi = \Sigma_1^{-1} \Delta_\mathcal{D}^2 \Sigma_1^{-1}$, $\Delta_\mathcal{D} = \text{diag}\left(\sigma_{j \in \mathcal{D}}^2\right)$ and $\sigma_j^2$ is defined in (2.6). Under Condition 1, $\sigma_j^2 \leq \tau_{\max}$ and $\lambda_{\max}\left(\Sigma_1^{-1}\right) \leq \frac{1}{\tau_{\min}}$, and $\lambda_{\max}\left(\Psi^T \Psi\right) \leq \left(\frac{\tau_{\max}}{\tau_{\min}}\right)^2$. Thus,

$$\text{trace}\left(\text{Var}\left(\alpha_\mathcal{D}(Y)\right)\right) = \text{trace}\left(\Psi \text{Var}\left(\alpha_{\mathcal{D}|\mathcal{B}}(Y)\right) \Psi^T\right) \leq \left(\frac{\tau_{\max}}{\tau_{\min}}\right)^2 \text{trace}\left(\text{Var}\left(\alpha_{\mathcal{D}|\mathcal{B}}(Y)\right)\right),$$

where $\alpha_{\mathcal{D}}(Y) = \alpha_{\mathcal{D}}^{(h)}$ and $\alpha_{\mathcal{D}|\mathcal{B}}(Y) = \alpha_{\mathcal{D}|\mathcal{B}}^{(h)}$ when $Y \in S_h$. Furthermore,

$$\text{trace}\left(\text{Var}\left(\alpha_{\mathcal{D}|\mathcal{B}}(Y)\right)\right) \geq |\mathcal{D}| \left(\frac{\tau_{\min}}{\tau_{\max}}\right)^2 \xi n^{-\kappa}.$$

Assume

$$\text{Cov}\left(\mathbf{X}_{\mathcal{E}\cup\mathcal{D}}|\mathbf{X}_{\mathcal{B}}\right) = \begin{pmatrix} \Sigma_0 & \Sigma_{01} \\ \Sigma_{10} & \Sigma_1 \end{pmatrix},$$

Under model (2.3), conditioning on $\mathbf{X}_{\mathcal{A}} = \mathbf{X}_{\mathcal{B}\cup\mathcal{D}}$, the distribution of $\mathbf{X}_{\mathcal{E}}$ does not depend on the slice. Therefore, the conditional distribution of $\mathbf{X}_{\mathcal{D}}$ given $\mathbf{X}_{\mathcal{C}} = \mathbf{X}_{\mathcal{B}\cup\mathcal{E}}$ can be written as

$$\mathbf{X}_{\mathcal{D}}|\mathbf{X}_{\mathcal{C}}, Y \in S_h \sim N\left(\alpha_{\mathcal{D}|\mathcal{C}}^{(h)} + \boldsymbol{\beta}_{\mathcal{D}|\mathcal{C}}^T \mathbf{X}_{\mathcal{C}}, \Sigma_{\mathcal{D}|\mathcal{C}}\right),$$

where $\alpha_{\mathcal{D}|\mathcal{C}}^{(h)} = \alpha_0 + \mathbf{M}\alpha_{\mathcal{D}|\mathcal{B}}^{(h)}$, $\mathbf{M} = \Sigma_1^{\frac{1}{2}}\left(I_{|\mathcal{D}|} - \mathbf{N}^T\left(I_{|\mathcal{E}|} + \mathbf{N}\mathbf{N}^T\right)^{-1}\mathbf{N}\right)\Sigma_1^{-\frac{1}{2}}$, $\mathbf{N} = \Sigma_0^{-\frac{1}{2}}\Sigma_{01}\Sigma_1^{-\frac{1}{2}}$ and $\alpha_0$ is a constant that does not depend on the slice. Since $\lambda_{\max}\left(\mathbf{N}\mathbf{N}^T\right) \leq \frac{\lambda_{\max}(\Sigma_0)}{\lambda_{\min}(\Sigma_0)} \leq \frac{\tau_{\max}}{\tau_{\min}}$, $\lambda_{\min}\left(\Sigma_1\right) \geq \tau_{\min}$ and $\lambda_{\min}\left(\Sigma_1^{-1}\right) \geq \frac{1}{\tau_{\max}}$, we have

$$\lambda_{\min}\left(I_{|\mathcal{D}|} - \mathbf{N}^T\left(I_{|\mathcal{E}|} + \mathbf{N}\mathbf{N}^T\right)^{-1}\mathbf{N}\right) \geq \frac{1}{1 + \frac{\tau_{\max}}{\tau_{\min}}} = \frac{\tau_{\min}}{\tau_{\max} + \tau_{\min}},$$

$$\begin{aligned}
& \text{trace}\left(\text{Var}\left(\alpha_{\mathcal{D}|\mathcal{C}}(Y)\right)\right) = \text{trace}\left(\mathbf{M}\text{Var}\left(\alpha_{\mathcal{D}|\mathcal{B}}(Y)\right)\mathbf{M}^T\right) \\
\geq\ & \lambda_{\min}\left(\Sigma_1^{-1}\right)\lambda_{\min}\left(\Sigma_1\right)\lambda_{\min}^2\left(I_{|\mathcal{D}|} - \mathbf{N}^T\left(I_{|\mathcal{E}|} + \mathbf{N}\mathbf{N}^T\right)^{-1}\mathbf{N}\right)\text{trace}\left(\text{Var}\left(\alpha_{\mathcal{D}|\mathcal{B}}(Y)\right)\right) \\
\geq\ & \left(\frac{\tau_{\min}}{\tau_{\max} + \tau_{\min}}\right)^2 \left(\frac{\tau_{\min}}{\tau_{\max}}\right)\left(\text{Var}\left(\alpha_{\mathcal{D}|\mathcal{B}}(Y)\right)\right) \\
\geq\ & |\mathcal{D}|\left(\frac{\tau_{\min}}{\tau_{\max} + \tau_{\min}}\right)^2 \left(\frac{\tau_{\min}}{\tau_{\max}}\right)^3 \xi n^{-\kappa}
\end{aligned}$$

Thus, there exists $j \in \mathcal{D} = \mathcal{C}^c \cap \mathcal{A}$ such that

$$\mathrm{Var}\left(\alpha_{j|\mathcal{C}}(Y)\right) \geq \left(\frac{\tau_{\min}}{\tau_{\max}+\tau_{\min}}\right)^2 \left(\frac{\tau_{\min}}{\tau_{\max}}\right)^3 \xi n^{-\kappa}.$$

For such $j$, we have $M_j = \alpha_{j|\mathcal{C}}(Y) + \boldsymbol{\beta}_{j|\mathcal{C}}^T \mathbf{X}_{\mathcal{C}}$, $V_j = \Sigma_{j|\mathcal{C}} \leq \tau_{\max}$, and

$$\begin{aligned} & \mathrm{Var}\left(M_j\right) - \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)^T \\ = \; & \mathrm{Var}\left(\alpha_{j|\mathcal{C}}(Y)\right) - \mathrm{Cov}\left(\alpha_{j|\mathcal{C}}(Y), \mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(\alpha_{j|\mathcal{C}}(Y), \mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)^T. \end{aligned}$$

Let $\mathbf{T}_1 = \mathrm{Cov}\left(\mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)$ and $\mathbf{T}_2 = \mathbb{E}\left(\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right) = \mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)$. Since $\lambda_{\min}\left(\mathbf{T}_2\right) \geq \tau_{\min}$ and $\lambda_{\max}\left(\mathbf{T}_1\right) \leq \lambda_{\max}\left(\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right) \leq \tau_{\max}$, $\lambda_{\min}\left(\mathbf{T}_1^{-\frac{1}{2}}\mathbf{T}_2\mathbf{T}_1^{-\frac{1}{2}}\right) \geq \frac{\tau_{\min}}{\tau_{\max}}$ and

$$\begin{aligned} & \mathrm{Cov}\left(\alpha_{j|\mathcal{C}}(Y), \mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(\alpha_{j|\mathcal{C}}(Y), \mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)^T \\ = \; & \mathrm{Cov}\left(\alpha_{j|\mathcal{C}}(Y), \mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)\left(\mathbf{T}_1+\mathbf{T}_2\right)^{-1}\mathrm{Cov}\left(\alpha_{j|\mathcal{C}}(Y), \mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)^T \\ \leq \; & \frac{1}{1+\lambda_{\min}\left(\mathbf{T}_1^{-\frac{1}{2}}\mathbf{T}_2\mathbf{T}_1^{-\frac{1}{2}}\right)}\mathrm{Cov}\left(\alpha_{j|\mathcal{C}}(Y), \mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)\mathbf{T}_1^{-1}\mathrm{Cov}\left(\alpha_{j|\mathcal{C}}(Y), \mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)^T \\ \leq \; & \frac{\tau_{\max}}{\tau_{\min}+\tau_{\max}}\mathrm{Cov}\left(\alpha_{j|\mathcal{C}}(Y), \mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)\left[\mathrm{Cov}\left(\mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)\right]^{-1}\mathrm{Cov}\left(\alpha_{j|\mathcal{C}}(Y), \mathbb{E}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right)^T \\ \leq \; & \frac{\tau_{\max}}{\tau_{\min}+\tau_{\max}}\mathrm{Var}\left(\alpha_{j|\mathcal{C}}(Y)\right). \end{aligned}$$

Therefore,

$$\begin{aligned} & \frac{\mathrm{Var}\left(M_j\right) - \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)^T}{V_j} \\ \geq \; & \frac{1}{\tau_{\max}}\frac{\tau_{\min}}{\tau_{\min}+\tau_{\max}}\mathrm{Var}\left(\alpha_{j|\mathcal{C}}(Y)\right) \\ \geq \; & \frac{1}{\tau_{\max}}\left(\frac{\tau_{\min}}{\tau_{\max}+\tau_{\min}}\right)^3\left(\frac{\tau_{\min}}{\tau_{\max}}\right)^3\xi n^{-\kappa}, \end{aligned}$$

where $\xi_1 = \frac{1}{\tau_{\max}} \left(\frac{\tau_{\min}}{\tau_{\max}+\tau_{\min}}\right)^3 \left(\frac{\tau_{\min}}{\tau_{\max}}\right)^3 \xi$. □

*Proof of Lemma 2.* Since $\mathbf{X}_{\mathcal{C}}$ follows either a multivariate normal distribution or a finite mixture of multivariate normal distributions, under Condition 1, one can show that for $i = 1, 2, \ldots, q$ and any $\epsilon > 0$ there exists constant $C_1$ and $C_2$ such that

$$\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \left|\widehat{\lambda}_i^{\mathcal{C}} - \lambda_i^{\mathcal{C}}\right| > \epsilon\right) \leq 2p(p+1)C_1 \exp\left(-C_2 n \frac{\tau_{\min}^2 \epsilon^2}{16p^2}\right)$$

following similar arguments in the proof of Lemma 2 in Zhong et al. (2012). Since $\Omega_{\mathcal{C}} = W_{\mathcal{C}} + B_{\mathcal{C}}$, where $\Omega_{\mathcal{C}} = \mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)$ and $W_{\mathcal{C}} = \mathbb{E}\left(\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)\right) = \mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}|S(Y)\right)$ under model (2.3), $\lambda_{\max}\left(\Omega_{\mathcal{C}}\right) \leq \tau_{\max}$ and $\lambda_{\min}\left(W_{\mathcal{C}}\right) \geq \tau_{\min}$ Thus,

$$\lambda_1^{\mathcal{C}} = \max_{||\boldsymbol{\eta}||=1} \frac{\boldsymbol{\eta}^T B_{\mathcal{C}} \boldsymbol{\eta}}{\boldsymbol{\eta}^T \Omega_{\mathcal{C}} \boldsymbol{\eta}} = 1 - \min_{||\boldsymbol{\eta}||=1} \frac{\boldsymbol{\eta}^T W_{\mathcal{C}} \boldsymbol{\eta}}{\boldsymbol{\eta}^T \Omega_{\mathcal{C}} \boldsymbol{\eta}} \leq 1 - \frac{\min_{||\boldsymbol{\eta}||=1} \boldsymbol{\eta}^T W_{\mathcal{C}} \boldsymbol{\eta}}{\max_{||\boldsymbol{\eta}||=1} \boldsymbol{\eta}^T \Omega_{\mathcal{C}} \boldsymbol{\eta}} \leq 1 - \frac{\tau_{\min}}{\tau_{\max}},$$

and

$$1 - \lambda_i^{\mathcal{C}} \geq 1 - \lambda_1 \geq \frac{\tau_{\min}}{\tau_{\max}}, \text{ for } i = 1, 2, \ldots, q.$$

Therefore, for $i = 1, 2, \ldots, q$ and $0 < \epsilon < 1$, there exists constant $C_1$ and $C_2$ such that

$$\begin{aligned} &\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \left|\log\left(1 - \widehat{\lambda}_i^{\mathcal{C}}\right) - \log\left(1 - \lambda_i^{\mathcal{C}}\right)\right| > \epsilon\right) \leq \Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \frac{\left|\widehat{\lambda}_i^{\mathcal{C}} - \lambda_i^{\mathcal{C}}\right|}{1 - \lambda_i^{\mathcal{C}}} > \frac{\epsilon}{2}\right) \\ \leq\ &\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \left|\widehat{\lambda}_i^{\mathcal{C}} - \lambda_i^{\mathcal{C}}\right| > \frac{\tau_{\min}}{2\tau_{\max}}\epsilon\right) \leq 2p(p+1)C_1 \exp\left(-C_2 n \frac{\tau_{\min}^4 \epsilon^2}{64\tau_{\max}^2 p^2}\right). \end{aligned}$$

as $n \to \infty$. □

*Proof of Theorem 1.* Let $R_{\mathcal{C}} = \sum_{i=1}^q \log\left(1 - \widehat{\lambda}_i^{\mathcal{C}}\right) - \sum_{i=1}^q \log\left(1 - \lambda_i^{\mathcal{C}}\right)$. Then, according to

Lemma 2, for $0 < \epsilon < 1$, there exists constant $C_1$ and $C_2$ such that

$$\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} |R_{\mathcal{C}}| > \epsilon\right) \leq 2p(p+1)qC_1 \exp\left(-C_2 n \frac{\tau_{\min}^4 \epsilon^2}{64\tau_{\max}^2 p^2 q^2}\right).$$

Under Condition 2, $p = o\left(n^{\rho}\right)$ with $2\rho + 2\kappa < 1$, and for any positive constant $C$,

$$\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} |R_{\mathcal{C}}| > Cn^{-\kappa}\right) \leq 2p(p+1)qC_1 \exp\left(-C_2 n^{1-2\kappa-2\rho} \frac{\tau_{\min}^4 C^2}{64\tau_{\max}^2 q^2}\right) \to 0$$

as $n \to \infty$. For $j \notin \mathcal{C}$ and $d = |\mathcal{C}|$,

$$\begin{aligned}
\widehat{D}_{j|\mathcal{C}} &= -\sum_{i=1}^{q} \log\left(1 - \widehat{\lambda}_i^{d+1}\right) + \sum_{i=1}^{q} \log\left(1 - \widehat{\lambda}_i^{d}\right) \\
&= -\sum_{i=1}^{q} \log\left(1 - \lambda_i^{d+1}\right) + \sum_{i=1}^{q} \log\left(1 - \lambda_i^{d}\right) - R_{[\mathcal{C}\cup\{j\}]} + R_{\mathcal{C}} \\
&= \log\left(1 + \frac{\mathrm{Var}\left(M_j\right) - \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)^T}{V_j}\right) \\
&\quad -R_{[\mathcal{C}\cup\{j\}]} + R_{\mathcal{C}},
\end{aligned}$$

where $M_j = \mathbb{E}\left(X_j|\mathbf{X}_{\mathcal{C}}, S(Y)\right)$, $V_j = \mathrm{Var}\left(X_j|\mathbf{X}_{\mathcal{C}}, S(Y)\right)$, and $V_j$ is a constant that does not depend on $\mathbf{X}_{\mathcal{C}}$ or $S(Y)$ under model (2.3).

When $\mathcal{C}^c \cap \mathcal{A} \neq \emptyset$, according to Lemma 1, there exist $\kappa > 0$ and $\xi_1 > 0$ such that

$$\max_{j\in\mathcal{C}^c\cap\mathcal{A}} \left[\frac{\mathrm{Var}\left(M_j\right) - \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)^T}{V_j}\right] \geq \xi_1 n^{-\kappa},$$

Then, for sufficiently large $n$, there exists $j \in \mathcal{C}^c \cap \mathcal{A}$ such that

$$\log\left(1 + \frac{\mathrm{Var}\left(M_j\right) - \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)^T}{V_j}\right) \geq \frac{\xi_1}{2} n^{-\kappa},$$

and

$$\widehat{D}_{j|\mathcal{C}} \geq \frac{\xi_1}{2} n^{-\kappa} - \left( \left| R_{[\mathcal{C}\cup\{j\}]} \right| + |R_{\mathcal{C}}| \right).$$

Let $c = \frac{\xi_1}{4}$. Since

$$\Pr\left( \max_{\mathcal{C}\subset\{1,2,\ldots,p\}} |R_{\mathcal{C}}| > \frac{c}{2} n^{-\kappa} \right) \to 0,$$

we have

$$\Pr\left( \min_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{A}\neq\emptyset} \max_{j\in\mathcal{C}^c\cap\mathcal{A}} \widehat{D}_{j|\mathcal{C}} \geq c n^{-\kappa} \right) \to 1,$$

as $n \to \infty$.

When variable $\mathcal{C}^c \cap \mathcal{A} = \emptyset$, for $j \in \mathcal{C}^c \subset \mathcal{A}^c$, $M_j = \mathbb{E}\left(X_j | \mathbf{X}_{\mathcal{C}}, S(Y)\right) = \mathbb{E}\left(X_j | \mathbf{X}_{\mathcal{C}}\right)$ is a linear combination of $\mathbf{X}_{\mathcal{C}}$ under model (2.3), and

$$\frac{\mathrm{Var}\left(M_j\right) - \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right) \left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1} \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)^T}{V_j} = 0.$$

Thus,

$$\widehat{D}_{j|\mathcal{C}} \leq \left( \left| R_{[\mathcal{C}\cup\{j\}]} \right| + |R_{\mathcal{C}}| \right),$$

and

$$\Pr\left( \max_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{A}=\emptyset} \max_{j\in\mathcal{C}^c} \widehat{D}_{j|\mathcal{C}} \geq C n^{-\kappa} \right) \leq \Pr\left( \max_{\mathcal{C}\subset\{1,2,\ldots,p\}} |R_{\mathcal{C}}| \geq \frac{C}{2} n^{-\kappa} \right) \to 0$$

for any positive constant $C$ as $n \to \infty$. □

## A.3 Proof of Proposition 2 in Section 2.1.2

We start with a simple case to get some intuitions behind the proof. Suppose there are two different sets of *minimally relevant* predictors $\{X_1\}$ and $\{X_2\}$. Under model (2.8)

and Condition 1, the support for the joint distribution of two predictors $X_1$ and $X_2$ is the entire 2-dimensional space $\mathbb{R}^2$. If $X_1 \perp\!\!\!\perp S(Y)|X_2$ and $X_2 \perp\!\!\!\perp S(Y)|X_1$, then we have $\Pr(S(Y)|X_1 = x, X_2 = x_2) = \Pr(S(Y)|X_2 = x_2) = \Pr(S(Y)|X_1 = x_1)$ for any $x_1, x_2 \in \mathbb{R}$. Thus, both $\Pr(S(Y)|X_2)$ and $\Pr(S(Y)|X_1)$ are constants, and $X_1 \perp\!\!\!\perp S(Y)$ and $X_2 \perp\!\!\!\perp S(Y)$, which is contradictory to the assumption that $\{X_1\}$ and $\{X_2\}$ are minimally relevant.

*Proof of Proposition 2.* Suppose there are two different sets of *minimally relevant* predictors indexed by $\mathcal{B}$ and $\mathcal{C}$, respectively. Define $\mathcal{A}_1 = \mathcal{B} \cap \mathcal{C}$, $\mathcal{A}_2 = \mathcal{B} \cap \mathcal{C}^c$ and $\mathcal{A}_3 = \mathcal{B}^c \cap \mathcal{C}$. Then, we must have $\mathcal{A}_2 \neq \emptyset$ and $\mathcal{A}_3 \neq \emptyset$. The conditional distribution of $\mathbf{X}_{\mathcal{A}_2}$ given $\mathbf{X}_{\mathcal{A}_1}$ and slice $S_h$ can be written as

$$\mathbf{X}_{\mathcal{A}_2}|\mathbf{X}_{\mathcal{A}_1}, Y \in S_h \sim N\left(\alpha^{(h)}_{\mathcal{A}_2|\mathcal{A}_1} + \mathbf{B}^{(h)}_{\mathcal{A}_2|\mathcal{A}_1}\mathbf{X}_{\mathcal{A}_1}, \Sigma_2^{(h)} = \Sigma^{(h)}_{\mathcal{A}_2|\mathcal{A}_1}\right).$$

Let $\alpha^{(h)} = \alpha^{(h)}_{\mathcal{A}_2|\mathcal{A}_1}$ and $\mathbf{B}^{(h)} = \mathbf{B}^{(h)}_{\mathcal{A}_2|\mathcal{A}_1} = \left(\mathbf{B}_j^{(h)}\right)_{j\in\mathcal{A}_1}$. Since $\mathcal{B}$ is minimally relevant, at least one of $\alpha^{(h)}$, $\mathbf{B}_j^{(h)}$ $(j \in \mathcal{A}_1)$ and $\Sigma_2^{(h)}$ has to be different across slices. The conditional distribution of $\mathbf{X}_{\mathcal{A}_3}$ given $\mathbf{X}_{\mathcal{A}_1\cup\mathcal{A}_2} = \mathbf{X}_{\mathcal{B}}$ is the same across slices,

$$\mathbf{X}_{\mathcal{A}_3}|\mathbf{X}_{\mathcal{B}}, Y \in S_h \sim N\left(a + \mathbf{b}\mathbf{X}_{\mathcal{A}_1} + \mathbf{c}\mathbf{X}_{\mathcal{A}_2}, \Sigma_3 = \Sigma_{\mathcal{B}}\right).$$

The conditional distribution of $\mathbf{X}_{\mathcal{A}_3}$ given $\mathbf{X}_{\mathcal{A}_1}$ and slice $S_h$ is

$$\mathbf{X}_{\mathcal{A}_3}|\mathbf{X}_{\mathcal{A}_1}, Y \in S_h \sim N\left(a + \mathbf{c}\alpha^{(h)} + \left(\mathbf{b} + \mathbf{c}\mathbf{B}^{(h)}\right)\mathbf{X}_{\mathcal{A}_1}, \Sigma_3 + \mathbf{c}\Sigma_2^{(h)}\mathbf{c}^T\right),$$

Since $\mathcal{A}_3 \in \mathcal{C}$ and $\mathcal{C}$ is minimally relevant, at least one of $\mathbf{c}\alpha^{(h)}$, $\mathbf{c}\mathbf{B}_j^{(h)}$ $(j \in \mathcal{A}_1)$ and $\mathbf{M}^{(h)} = \mathbf{c}\Sigma_2^{(h)}\mathbf{c}^T$ has to be different across slices. Given $\mathbf{X}_{\mathcal{A}_1\cup\mathcal{A}_3} = \mathbf{X}_{\mathcal{C}}$ and slice $S_h$, the conditional

distribution of $\mathbf{cX}_{\mathcal{A}_2}$ is

$$\mathbf{cX}_{\mathcal{A}_2}|\mathbf{X}_{\mathcal{C}}, Y \in S_h \sim N\left(\gamma^{(h)} + \mathbf{D}^{(h)}\mathbf{X}_{\mathcal{A}_1} + \mathbf{C}^{(h)}\mathbf{X}_{\mathcal{A}_3}, \Sigma^{(h)}\right),$$

where $\mathbf{C}^{(h)} = \mathbf{M}^{(h)}\left(\Sigma_3 + \mathbf{M}^{(h)}\right)^{-1}$, $\gamma^{(h)} = -\mathbf{C}^{(h)}a + \left(I_{|\mathcal{A}_3|} - \mathbf{C}^{(h)}\right)\mathbf{c}\alpha^{(h)}$, $\mathbf{D}^{(h)} = -\mathbf{C}^{(h)}\mathbf{b} + \left(I_{|\mathcal{A}_3|} - \mathbf{C}^{(h)}\right)\mathbf{cB}^{(h)}$, and $\Sigma^{(h)} = \mathbf{M}^{(h)} - \mathbf{M}^{(h)}\left(\Sigma_3 + \mathbf{M}^{(h)}\right)^{-1}\mathbf{M}^{(h)}$.

First, if $\mathbf{M}^{(h)}$ $(h = 1, 2, \ldots, H)$ are different across slices. Then $\mathbf{N}^{(h)} = \Sigma_3^{-\frac{1}{2}}\mathbf{M}^{(h)}\Sigma_3^{-\frac{1}{2}} = \Gamma^{(h)}\Lambda^{(h)}\left[\Gamma^h\right]^{-1}$ are different across slices (note that under Condition 1, $\Sigma_3^{\frac{1}{2}}$ is invertible). We have

$$\Sigma_3^{-\frac{1}{2}}\Sigma^{(h)}\Sigma_3^{-\frac{1}{2}} = \mathbf{N}^{(h)} - \mathbf{N}^{(h)}\left(I_{|\mathcal{A}_3|} + \mathbf{N}^{(h)}\right)^{-1}\mathbf{N}^{(h)} = \Gamma^{(h)}\Lambda^{(h)}\left(I_{|\mathcal{A}_3|} + \Lambda^{(h)}\right)^{-1}\left[\Gamma^h\right]^{-1}.$$

Thus, $\Sigma_3^{-\frac{1}{2}}\Sigma^{(h)}\Sigma_3^{-\frac{1}{2}}$ and $\Sigma^{(h)}$ are different across slices.

Second, if $\mathbf{M}^{(h)} = \mathbf{M}$ $(h = 1, 2, \ldots, H)$ are the same across slices. Then at least one of $\mathbf{c}\alpha^{(h)}$ and $\mathbf{cB}_j^{(h)}$ $(j \in \mathcal{A}_1)$ has to be different across slices. Without loss of generality, assume $\mathbf{c}\alpha^{(h)}$ are different across slices, that is, $\text{trace}\left(\text{Var}\left(\mathbf{c}\alpha(Y)\right)\right) > 0$, where $\alpha(Y) = \alpha^{(h)}$ when $Y \in S_h$. Under Condition 1, $\lambda_{\max}\left(\mathbf{N}\right) \leq \frac{\tau_{\max}}{\tau_{\min}}$, $\lambda_{\min}\left(\Sigma_3\right) \geq \tau_{\min}$ and $\lambda_{\min}\left(\Sigma_3^{-1}\right) \geq \frac{1}{\tau_{\max}}$. Then,

$$\mathbf{C}^{(h)} = \mathbf{C} = I_{\mathcal{A}_3} - \mathbf{M}\left(\Sigma_3 + \mathbf{M}\right)^{-1} = \Sigma_3^{\frac{1}{2}}\left(I_{\mathcal{A}_3} - \mathbf{N}\left(I_{\mathcal{A}_3} + \mathbf{N}\right)^{-1}\right)\Sigma_3^{-\frac{1}{2}},$$

and

$$\begin{aligned}
& \text{trace}\left(\text{Var}\left(\gamma(Y)\right)\right) = \text{trace}\left(\mathbf{C}\text{Var}\left(c\alpha(Y)\right)\mathbf{C}^T\right) \\
\geq\ & \lambda_{\min}\left(\Sigma_3^{-1}\right)\lambda_{\min}\left(\Sigma_3\right)\lambda_{\min}^2\left(I_{\mathcal{A}_3} - \mathbf{N}\left(I_{\mathcal{A}_3} + \mathbf{N}\right)^{-1}\right)\text{trace}\left(\text{Var}\left(\mathbf{c}\alpha(Y)\right)\right) \\
\geq\ & \left(\frac{\tau_{\max}}{\tau_{\max} + \tau_{\min}}\right)^2\left(\frac{\tau_{\max}}{\tau_{\min}}\right)\text{trace}\left(\text{Var}\left(\mathbf{c}\alpha(Y)\right)\right) > 0,
\end{aligned}$$

where $\gamma(Y) = \gamma^{(h)}$ when $Y \in S_h$. Thus, $\gamma^{(h)}$ are different across slices.

Therefore, given $\mathbf{X}_{\mathcal{C}}$ and slice $S_h$, the conditional distribution of $\mathbf{c}\mathbf{X}_{\mathcal{A}_2}$ depends on slice $S_h$, which is contradictory with the previous assumption that $\mathcal{A}_2 \in \mathcal{C}^c$ and the conditional distribution of $\mathbf{X}_{\mathcal{A}_2}$ given $\mathbf{X}_{\mathcal{C}}$ is the same across slices. So we must have $\mathcal{B} = \mathcal{C}$. □

## A.4 Proof of Properties of Augmented Test Statistic in Section 2.1.2

Properties of the likelihood-ratio test statistic proposed in Section 2.1.2 are summarized in the following proposition:

**Proposition 5.** *Given the current set of selected predictors indexed by $\mathcal{C}$ with dimension $|\mathcal{C}| = d$ and another predictor indexed by $j \notin \mathcal{C}$, the scaled log-likelihood-ratio test statistic for testing*

$$H_0 : \mathcal{A} = \mathcal{C} \ \textit{v.s.} \ H_1 : \mathcal{A} = \mathcal{C} \cup \{j\},$$

*under the augmented model (2.8) can be written as*

$$\widehat{D}^*_{j|\mathcal{C}} = \log \widehat{\sigma}^2_j - \sum_{h=1}^{H} \frac{n_h}{n} \log \left[\widehat{\sigma}^{(h)}_j\right]^2, \tag{A.5}$$

*where $\left[\widehat{\sigma}^{(h)}_j\right]^2$ is the estimated variance by regressing $X_j$ on $\mathbf{X}_{\mathcal{C}}$ in slice $S_h$, and $\widehat{\sigma}^2_j$ is the estimated variance by regressing $X_j$ on $\mathbf{X}_{\mathcal{C}}$ using all the observations.*

*(a) For any fixed slicing scheme and $\mathcal{A} \subset \mathcal{C}$,*

$$\widehat{D}^*_{j|\mathcal{C}} \sim \log\left(1 + \frac{Q_0}{\sum_{h=1}^{H} Q_h}\right) - \sum_{h=1}^{H} \frac{n_h}{n} \log\left(\frac{Q_h/n_h}{\sum_{h=1}^{H} Q_h/n}\right) \xrightarrow{D} \chi^2_{(H-1)(d+2)},$$

*where* $Q_0 \sim \chi^2_{(H-1)(d+1)}$ *and* $Q_h \sim \chi^2_{n_h-(d+1)}$ $(1 \le h \le H)$ *are mutually independent.*

*(b) Under the same condition as in (a),*

$$\left(n\widehat{D}^*_{j|\mathcal{C}}\right)_{j\in\mathcal{C}^c} \xrightarrow{D} \left(\sum_{i=1}^{(H-1)(d+1)} z_{ij}^2 + \sum_{i=1}^{H-1} \widetilde{z}_{ij}^2\right)_{j\in\mathcal{C}^c},$$

*where* $\mathbf{z}_i$*'s and* $\widetilde{\mathbf{z}}_i$*'s are mutually independent with*

$$\mathbf{z}_i = (z_{ij})_{j\in\mathcal{C}^c} \sim MVN\left(\mathbf{0}, \left[Corr\left(X_j, X_k|\mathbf{X}_\mathcal{C}\right)\right]_{j,k\in\mathcal{C}^c}\right),$$

*and*

$$\widetilde{\mathbf{z}}_i = (\widetilde{z}_{ij})_{j\in\mathcal{C}^c} \sim MVN\left(\mathbf{0}, \left[Corr^2\left(X_j, X_k|\mathbf{X}_\mathcal{C}\right)\right]_{j,k\in\mathcal{C}^c}\right).$$

*(c) For any fixed slicing scheme, as* $n \to \infty$*,*

$$\begin{aligned} & \widehat{D}^*_{j|\mathcal{C}} \xrightarrow{a.s.} D^*_{j|\mathcal{C}} \\ = \; & \log\left(1 + \frac{Var(M_j) - Cov\left(M_j, \mathbf{X}_\mathcal{C}\right)\left[Cov(\mathbf{X}_\mathcal{C})\right]^{-1} Cov\left(M_j, \mathbf{X}_\mathcal{C}\right)^T}{\mathbb{E}\left(V_j\right)}\right) \\ & + \log \mathbb{E}\left(V_j\right) - \mathbb{E}\log\left(V_j\right) \end{aligned}$$

*where* $M_j = \mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}, S(Y)\right)$*,* $V_j = Var\left(X_j|\mathbf{X}_\mathcal{C}, S(Y)\right)$ *and* $S(Y) = h$ *when* $Y \in S_h$ *(*$1 \le h \le H$*). Furthermore,*

$$D^*_{j|\mathcal{C}} = 0 \text{ iff } \mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}, Y \in S_h\right) = \mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}\right), \text{ and } Var\left(X_j|\mathbf{X}_\mathcal{C}, Y \in S_h\right) = Var\left(X_j|\mathbf{X}_\mathcal{C}\right),$$

*for* $1 \le h \le H$*.*

*Proof of Proposition 5.* (a) Let $\widetilde{\mathbf{x}}_\mathcal{C} = [\mathbf{1}_n, \mathbf{x}_\mathcal{C}]$, where $\mathbf{1}_n$ is a $n$-dimensional column vector of

1 and $I_n$ is a $n$ by $n$ identity matrix. We denote $P_0 = \widetilde{\mathbf{x}}_{\mathcal{C}} \left(\widetilde{\mathbf{x}}_{\mathcal{C}}^T \widetilde{\mathbf{x}}_{\mathcal{C}}\right)^{-1} \widetilde{\mathbf{x}}_{\mathcal{C}}^T$,

$$P_h = diag\left(0, \ldots, \widetilde{\mathbf{x}}_{\mathcal{C}}^{(h)} \left(\widetilde{\mathbf{x}}_{\mathcal{C}}^{(h)T} \widetilde{\mathbf{x}}_{\mathcal{C}}^{(h)}\right)^{-1} \widetilde{\mathbf{x}}_{\mathcal{C}}^{(h)T}, \ldots, 0\right),$$

$R_h = diag(0, \ldots, I_{n_h}, \ldots, 0) - P_h$ and $R_0 = I_n - P_0$. Then,

$$\left[\widehat{\sigma}_j^{(h)}\right]^2 = \frac{1}{n_h} \mathbf{x}_j^T R_h \mathbf{x}_j = \frac{1}{n_h} \sigma_j^2 Q_h = \frac{1}{n} \frac{\sigma_j^2 Q_h}{s_h}, \text{ for } h = 1, 2, \ldots, H$$

and

$$\begin{aligned} \widehat{\sigma}_j^2 &= \frac{1}{n} \mathbf{x}_j^T R_0 \mathbf{x}_j = \frac{1}{n} \mathbf{x}_j^T \left(\sum_{h=1}^{H} R_h\right) \mathbf{x}_j + \frac{1}{n} \mathbf{x}_j^T \left(\sum_{h=1}^{H} P_h - P_0\right) \mathbf{x}_j \\ &= \frac{1}{n} \sum_{h=1}^{H} \sigma_j^2 Q_h + \frac{1}{n} \sigma_j^2 Q_0, \end{aligned}$$

where $\sigma_j^2$ is the conditional variance of $X_j$ given $\mathbf{X}_{\mathcal{C}}$ for $j \in \mathcal{C}^c$ and $\mathcal{A} \subset \mathcal{C}$.

The augmented likelihood-ratio test statistic in (A.5) can be written as

$$\begin{aligned} \widehat{D}_{j|\mathcal{C}}^* &= -\left(\sum_{h=1}^{H} s_h \log\left[\widehat{\sigma}_j^{(h)}\right]^2 - \log \widehat{\sigma}_j^2\right) \\ &= \left(\log\left(\frac{1}{n} \sum_{h=1}^{H} \sigma_j^2 Q_h + \frac{1}{n} \sigma_j^2 Q_0\right) - \sum_{h=1}^{H} s_h \log\left(\frac{1}{n} \frac{\sigma_j^2 Q_h}{s_h}\right)\right) \\ &= \left(\log\left(1 + \frac{Q_0}{\sum_{h=1}^{H} Q_h}\right) - \sum_{h=1}^{H} s_h \log\left(\frac{Q_h / s_h}{\sum_{h=1}^{H} Q_h}\right)\right). \end{aligned}$$

Note that both $\left(\sum_{h=1}^{H} P_h - P_0\right)$ and $R_h$'s are orthogonal to $\widetilde{\mathbf{x}}_{\mathcal{C}}$. Given that $j \in \mathcal{C}^c$, according to *Cochran's theorem*, $Q_0 = \mathbf{x}_j^T \left(\sum_{h=1}^{H} P_h - P_0\right) \mathbf{x}_j / \sigma_j^2$ and $Q_h = \mathbf{x}_j^T R_h \mathbf{x}_j / \sigma_j^2$ are

independent, and

$$Q_0 \sim \chi^2_{(H-1)(d+1)}, \text{ and } Q_h \sim \chi^2_{n_h-(d+1)}, \text{ for } h = 1, 2, \ldots, H, \text{ and } d = |\mathcal{C}|.$$

Let $n'_h = n_h - (d+1)$ and $n' = n - H(d+1)$. For any fixed slicing scheme, as $n \to \infty$, $n'_h \to \infty$, $n' \to \infty$ and $n'_h/n_h \to 1$, $n'/n \to 1$ given that $d/n \to 0$. Since

$$\frac{Q_0}{n'} \xrightarrow{P} 0, \ \frac{\sum_{h=1}^H Q_h}{n'} \xrightarrow{P} 1, \text{ and } \frac{Q_0}{\sum_{h=1}^H Q_h} = \frac{Q_0/n'}{\sum_{h=1}^H Q_h/n'} \xrightarrow{P} 0,$$

we have

$$U_j \equiv n \log\left(1 + \frac{Q_0}{\sum_{h=1}^H Q_h}\right) \xrightarrow{P} \frac{n}{n'} \frac{Q_0}{\sum_{h=1}^H Q_h/n'} \xrightarrow{P} Q_0 \sim \chi^2_{(H-1)(d+1)}.$$

Since

$$\frac{Q_h/s_h}{\sum_{h=1}^H Q_h} = \frac{n'_h/n_h}{n'/n} \frac{Q_h/n'_h}{\sum_{h=1}^H Q_h/n'} \xrightarrow{P} 1, \text{ for } h = 1, 2, \ldots, H,$$

and

$$\frac{\sum_{h=1}^H Q_h}{n} = \frac{n'}{n} \frac{\sum_{h=1}^H Q_h}{n'} \xrightarrow{P} 1,$$

we have

$$\begin{aligned}
\widetilde{U}_j &\equiv -n\sum_{h=1}^{H} s_h \log\left(\frac{Q_h/s_h}{\sum_{h=1}^{H} Q_h}\right) = -n\sum_{h=1}^{H} s_h \log\left(1+\frac{Q_h/s_h}{\sum_{h=1}^{H} Q_h}-1\right) \\
&\simeq -n\sum_{h=1}^{H} s_h\left(\frac{Q_h/s_h}{\sum_{h=1}^{H} Q_h}-1\right)+\frac{1}{2}\sum_{h=1}^{H} n_h\left(\frac{Q_h/s_h}{\sum_{h=1}^{H} Q_h}-1\right)^2 \\
&= \frac{1}{2}\sum_{h=1}^{H} n_h\left(\frac{Q_h/n_h}{\sum_{h=1}^{H} Q_h/n}-1\right)^2 = \frac{1}{2}\frac{\sum_{h=1}^{H} n_h\left(Q_h/n_h-\sum_{h=1}^{H} Q_h/n\right)^2}{\left(\sum_{h=1}^{H} Q_h/n\right)^2} \\
&\simeq \frac{1}{2}\sum_{h=1}^{H} n_h\left(Q_h/n_h-\sum_{h=1}^{H} Q_h/n\right)^2 = \sum_{h=1}^{H}\left(\frac{Q_h}{\sqrt{2n_h}}\right)^2-\left(\sum_{h=1}^{H}\sqrt{\frac{n_h}{n}}\frac{Q_h}{\sqrt{2n_h}}\right)^2.
\end{aligned}$$

Let $q_h = \frac{Q_h}{\sqrt{2n_h}}$, for $h = 1, 2, \ldots, H$. Then

$$q_h = \sqrt{\frac{n'_h}{n_h}}\frac{\sum_{i=1}^{n'_h}\left[z_{ij}^{(h)}\right]^2}{\sqrt{2n'_h}},$$

where $z_{ij}^{(h)} \sim N(0,1)$ independently for $i = 1, 2, \ldots, n'_h$ and $h = 1, 2, \ldots, H$. Thus, according to central limit theorem, $q_h \xrightarrow{D} N(0,1)$, for $h = 1, 2, \ldots, H$, and let $\mathbf{q} = (q_1, \ldots, q_H)^T$. Then,

$$\widetilde{U}_j \simeq \sum_{h=1}^{H} q_h^2 - \left(\sum_{h=1}^{H}\sqrt{\frac{n_h}{n}}q_h\right)^2 = \mathbf{q}^T\left(I_H - JJ^T\right)\mathbf{q}$$

where $J = \left(\sqrt{n_1/n}, \ldots, \sqrt{n_H/n}\right)^T$ and $J^TJ = 1$. According to Cochran's theorem, $\widetilde{U}_j$ is asymptotically $\chi^2_{H-1}$. Since $Q_0$ is independent of $Q_h$ $(h = 1, 2, \ldots, H)$, $U_j$ is asymptotically independent of $\widetilde{U}_j$ and

$$n\widehat{D}^*_{j|\mathcal{C}} = U_j + \widetilde{U}_j \xrightarrow{D} \chi^2_{(H-1)(d+2)}.$$

To prove (b), for any $j \in \mathcal{C}^c$, we denote

$$U_j \simeq Q_j^{(0)} = \frac{\mathbf{x}_j^T \left( \sum_{h=1}^{H} P_h - P_0 \right) \mathbf{x}_j}{\sigma_j^2} = \sum_{i=1}^{(H-1)(d+1)} z_{ij}^2,$$

where $z_{ij} \sim N(0,1)$ independently for the same $j$ and $i = 1, 2, \ldots, (H-1)(d+1)$, and $\mathrm{Cov}(z_{ij}, z_{ij'}) = \mathrm{Corr}\left(X_j, X_{j'} | \mathbf{X}_{\mathcal{C}}\right)$ for $j' \neq j$. For any $j$ and $h \in \{1, 2, \ldots, H\}$, we denote

$$Q_j^{(h)} = \frac{\mathbf{x}_j^T R_h \mathbf{x}_j}{\sigma_j^2} = \sum_{i=1}^{n_h'} \left[ z_{ij}^{(h)} \right]^2, \text{ and } q_j^{(h)} = \frac{Q_j^{(h)}}{\sqrt{2n_h}},$$

where $z_{ij}^{(h)} \sim N(0,1)$ independently for the same $j$ and $i = 1, 2, \ldots, n_h'$, and $\mathrm{Cov}(z_{ij}^{(h)}, z_{ij'}^{(h)}) = \mathrm{Corr}\left(X_j, X_{j'} | \mathbf{X}_{\mathcal{C}}\right)$ for $j' \neq j$. Since $\mathrm{Cov}\left( \left[ z_{ij}^{(h)} \right]^2, \left[ z_{ij'}^{(h)} \right]^2 \right) = 2\mathrm{Corr}^2\left(X_j, X_{j'} | \mathbf{X}_{\mathcal{C}}\right)$, for $h = 1, 2, \ldots . H$ and $j \neq j'$, we have

$$\mathrm{Cov}\left( q_j^{(h)}, q_{j'}^{(h)} \right) = \frac{1}{2n_h} \mathrm{Cov}\left( Q_j^{(h)}, Q_{j'}^{(h)} \right) = \frac{n_h'}{n_h} \mathrm{Corr}^2\left(X_j, X_{j'} | \mathbf{X}_{\mathcal{C}}\right) \to \mathrm{Corr}^2\left(X_j, X_{j'} | \mathbf{X}_{\mathcal{C}}\right).$$

Hence

$$\widetilde{U}_j \simeq \sum_{h=1}^{H} \left[ q_j^{(h)} \right]^2 - \left( \sum_{h=1}^{H} \sqrt{\frac{n_h}{n}} \left[ q_j^{(h)} \right] \right)^2 \simeq \sum_{i=1}^{(H-1)} \tilde{z}_{ij}^2,$$

where $\tilde{z}_{ij} \sim N(0,1)$ independent for the same $j$ and $i = 1, 2, \ldots, (H-1)$, and for $j' \neq j$, $\mathrm{Cov}(\tilde{z}_{ij}, \tilde{z}_{ij'}) = \mathrm{Corr}^2\left(X_j, X_{j'} | \mathbf{X}_{\mathcal{C}}\right)$. Therefore,

$$\left( n \widehat{D}_{j|\mathcal{C}}^* \right)_{j \in \mathcal{C}^c} = \left( U_j + \widetilde{U}_j \right)_{j \in \mathcal{C}^c} \xrightarrow{D} \left( \sum_{i=1}^{(H-1)(d+1)} z_{ij}^2 + \sum_{i=1}^{(H-1)} \tilde{z}_{ij}^2 \right)_{j \in \mathcal{C}^c}.$$

(c) For any fixed slicing scheme, as $n \to \infty$, $n_h \to \infty$ for $h = 1, 2, \ldots, H$. Under the

normality assumption, as $n_h \to \infty$,

$$\begin{aligned}
&\left[\widehat{\sigma}_j^{(h)}\right]^2 \xrightarrow{P} \left[\sigma_j^{(h)}\right]^2 \\
= \;& \mathrm{Var}\left(X_j | Y \in S_h\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} | Y \in S_h\right) \left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}} | Y \in S_h\right)\right]^{-1} \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} | Y \in S_h\right)^T \\
= \;& \mathrm{Var}\left(X_j | \mathbf{X}_{\mathcal{C}}, Y \in S_h\right),
\end{aligned}$$

and as $n \to \infty$,

$$\begin{aligned}
&\widehat{\sigma}_j^2 \xrightarrow{P} \sigma_j^2 \\
= \;& \mathrm{Var}\left(X_j\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right) \left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1} \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)^T \\
= \;& \mathrm{Var}\left(X_j\right) - \mathrm{Cov}\left(\mathbb{E}\left(X_j | \mathbf{X}_{\mathcal{C}}, S(Y)\right), \mathbf{X}_{\mathcal{C}}\right) \left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1} \mathrm{Cov}\left(\mathbb{E}\left(X_j | \mathbf{X}_{\mathcal{C}}, S(Y)\right), \mathbf{X}_{\mathcal{C}}\right)^T.
\end{aligned}$$

Let $M_j = \mathbb{E}\left(X_j | \mathbf{X}_{\mathcal{C}}, S(Y)\right)$ and $V_j = \mathrm{Var}\left(X_j | \mathbf{X}_{\mathcal{C}}, S(Y)\right)$. Then, $\mathrm{Var}\left(X_j\right) = \mathrm{Var}\left(M_j\right) + \mathbb{E}\left(V_j\right)$, and

$$\begin{aligned}
\widehat{D}^*_{j|\mathcal{C}} &= \log \widehat{\sigma}_j^2 - \sum_{h=1}^{H} s_h \log \left[\widehat{\sigma}_j^{(h)}\right]^2 \\
&\xrightarrow{P} \log \sigma_j^2 - \sum_{h=1}^{H} s_h \log \left[\sigma_j^{(h)}\right]^2 \\
&= \log\left(\mathbb{E}\left(V_j\right) + \mathrm{Var}\left(M_j\right) - \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right) \left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1} \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)^T\right) - \\
&\quad \sum_{h=1}^{H} s_h \log\left(\mathrm{Var}\left(X_j | \mathbf{X}_{\mathcal{C}}, Y \in S_h\right)\right).
\end{aligned}$$

Since $\mathrm{Var}\left(X_j | \mathbf{X}_{\mathcal{C}}, Y \in S_h\right)$ is a constant that does not depend on $\mathbf{X}_{\mathcal{C}}$ under the normality assumption,

$$\mathbb{E} \log(V_j) = \sum_{h=1}^{H} s_h \log\left(\mathrm{Var}\left(X_j | \mathbf{X}_{\mathcal{C}}, Y \in S_h\right)\right),$$

and thus

$$\begin{aligned}\widehat{D}^{*}_{j|\mathcal{C}} \xrightarrow{P} & \log\left(\mathbb{E}\left(V_j\right)+\mathrm{Var}\left(M_j\right)-\mathrm{Cov}\left(M_j,\mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(M_j,\mathbf{X}_{\mathcal{C}}\right)^{T}\right)\\ & -\mathbb{E}\log(V_j)\\ = & \log\left(1+\frac{\mathrm{Var}\left(M_j\right)-\mathrm{Cov}\left(M_j,\mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(M_j,\mathbf{X}_{\mathcal{C}}\right)^{T}}{\mathbb{E}\left(V_j\right)}\right)\\ & +\log\left(\mathbb{E}V_j\right)-\mathbb{E}\log\left(V_j\right).\end{aligned}$$

Note that

$$\mathrm{Var}\left(M_j\right)\geq\mathrm{Cov}\left(M_j,\mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(M_j,\mathbf{X}_{\mathcal{C}}\right)^{T}$$

where equality holds if and only if $M_j=\mathbb{E}\left(X_j|\mathbf{X}_{\mathcal{C}},S(Y)\right)$ is a linear combination of $\mathbf{X}_{\mathcal{C}}$ that does not depend on $S(Y)$, that is, $M_j=\mathbb{E}\left(X_j|\mathbf{X}_{\mathcal{C}},S(Y)\right)=\mathbb{E}\left(X_j|\mathbf{X}_{\mathcal{C}}\right)$ under the normality assumption. Furthermore, according to Jensen's inequality

$$\log\left(\mathbb{E}V_j\right)\geq\mathbb{E}\log\left(V_j\right),$$

where equality holds if and only if $V_j=\mathbb{E}V_j$, or equivalently, $\mathrm{Var}\left(X_j|\mathbf{X}_{\mathcal{C}},Y\in S_h\right)$ is a constant for $h=1,2,\ldots,H$. Combined with $M_j=\mathbb{E}\left(X_j|\mathbf{X}_{\mathcal{C}}\right)$, $\mathrm{Var}\left(X_j|\mathbf{X}_{\mathcal{C}}\right)=\mathbb{E}\left(V_j|\mathbf{X}_{\mathcal{C}}\right)+\mathrm{Var}\left(M_j|\mathbf{X}_{\mathcal{C}}\right)=V_j=\mathrm{Var}\left(X_j|\mathbf{X}_{\mathcal{C}},S(Y)\right)$. □

## A.5 Proof of Theorem 2 in Section 2.1.2

To prove Theorem 2, we will need the following lemma.

**Lemma 3.** *Under the same condition as in Theorem 2, for* $0<\epsilon<1$*, there exists positive*

*constants $C_1$ and $C_2$ such that*

$$Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}\max_{j\in\mathcal{C}^c}\left|\log\widehat{\sigma}_j^2-\log\sigma_j^2\right|>\epsilon\right)\leq\frac{p(p+1)}{2}C_1\exp\left(-C_2n\frac{\epsilon^2}{p^2L^2}\right),$$

*and*

$$\begin{aligned}&Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}\max_{j\in\mathcal{C}^c}\left|\sum_{h=1}^{H}s_h\log\left[\widehat{\sigma}_j^{(h)}\right]^2-\sum_{h=1}^{H}s_h\log\left[\sigma_j^{(h)}\right]^2\right|>\epsilon\right)\\ \leq\ &\frac{Hp(p+1)}{2}C_1\exp\left(-C_2n\frac{\epsilon^2}{H^2p^2L^2}\right),\end{aligned}$$

*where* $L=\frac{4}{\tau_{min}}\left(3\left(\frac{\tau_{max}}{\tau_{min}}\right)^{3/2}+1\right)$.

*Proof of Lemma 3.* We denote $V_{\mathcal{C}}=\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)=\left(v_{j_1,j_2}\right)_{j_1,j_2\in\mathcal{C}}$, $r_{j,\mathcal{C}}=\mathrm{Cov}\left(X_j,\mathbf{X}_{\mathcal{C}}\right)^T$ and $v_{j,j}=\mathrm{Var}\left(X_j\right)$, and $\widehat{V}_{\mathcal{C}}=\left(\widehat{v}_{j_1,j_2}\right)_{j_1,j_2\in\mathcal{C}}$, $\widehat{r}_{j,\mathcal{C}}$ and $\widehat{v}_{j,j}$ are the corresponding sample estimates. Then, $\sigma_j^2=v_{j,j}-r_{j,\mathcal{C}}^TV_{\mathcal{C}}^{-1}r_{j,\mathcal{C}}$, $\widehat{\sigma}_j^2=\widehat{v}_{j,j}-\widehat{r}_{j,\mathcal{C}}^T\widehat{V}_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}$, and

$$\left|\widehat{\sigma}_j^2-\sigma_j^2\right|\leq\left|\widehat{v}_{j,j}-v_{j,j}\right|+\left|\widehat{r}_{j,\mathcal{C}}^T\widehat{V}_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}-\widehat{r}_{j,\mathcal{C}}^TV_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}\right|+\left|\widehat{r}_{j,\mathcal{C}}^TV_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}-r_{j,\mathcal{C}}^TV_{\mathcal{C}}^{-1}r_{j,\mathcal{C}}\right|$$

Let $M=\max_{j_1,j_2\in\{1,2,\ldots,p\}}\left|\widehat{v}_{j_1,j_2}-v_{j_1,j_2}\right|$. We have $\max_{j\in\{1,2,\ldots,p\}}\left|\widehat{v}_{j,j}-v_{j,j}\right|\leq M$,

$$\left|\widehat{r}_{j,\mathcal{C}}^TV_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}-r_{j,\mathcal{C}}^TV_{\mathcal{C}}^{-1}r_{j,\mathcal{C}}\right|=\left|\left(\widehat{r}_{j,\mathcal{C}}-r_{j,\mathcal{C}}\right)^TV_{\mathcal{C}}^{-1}\left(\widehat{r}_{j,\mathcal{C}}+r_{j,\mathcal{C}}\right)\right|,$$

and

$$\begin{aligned}\left|\widehat{r}_{j,\mathcal{C}}^T\widehat{V}_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}-\widehat{r}_{j,\mathcal{C}}^TV_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}\right|&=\left|\left(\widehat{V}_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}\right)^T\left(\widehat{V}_{\mathcal{C}}-V_{\mathcal{C}}\right)\left(V_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}\right)\right|\\&=||\boldsymbol{\eta}_1||||\boldsymbol{\eta}_2||\left|\boldsymbol{\eta}_1^{*T}\left(\widehat{V}_{\mathcal{C}}-V_{\mathcal{C}}\right)\boldsymbol{\eta}_2^*\right|,\end{aligned}$$

where $\boldsymbol{\eta}_1 = \widehat{V}_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}$, $\boldsymbol{\eta}_2 = V_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}$, and $\boldsymbol{\eta}_1^* = \frac{\boldsymbol{\eta}_1}{||\boldsymbol{\eta}_1||}$, $\boldsymbol{\eta}_2^* = \frac{\boldsymbol{\eta}_2}{||\boldsymbol{\eta}_2||}$.

Since

$$\tau_{\min} \leq \min_{\mathcal{C}\subset\{1,2,\ldots,p\}} \lambda_{\min}\{V_{\mathcal{C}}\} \leq \max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \lambda_{\max}\{V_{\mathcal{C}}\} \leq \tau_{\max},$$

using similar arguments to the proof of Lemma 1 in Wang (2009) with $p = o(n^{\rho})$ and $2\rho + 2\kappa < 1$, we have

$$2^{-1}\tau_{\min} \leq \min_{\mathcal{C}\subset\{1,2,\ldots,p\}} \lambda_{\min}\{\widehat{V}_{\mathcal{C}}\} \leq \max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \lambda_{\max}\{\widehat{V}_{\mathcal{C}}\} \leq 2\tau_{\max}.$$

Then, $||\widehat{V}_{\mathcal{C}}^{-1/2}\widehat{r}_{j,\mathcal{C}}||^2 = \widehat{r}_{j,\mathcal{C}}^T\widehat{V}_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}} \leq \widehat{v}_{j,j} \leq 2\tau_{\max}$, $||\widehat{V}_{\mathcal{C}}^{1/2}|| \leq \sqrt{2\tau_{\max}}$, and $||\widehat{V}_{\mathcal{C}}^{-1/2}|| \leq \sqrt{\frac{2}{\tau_{\min}}}$. Thus,

$$||\boldsymbol{\eta}_1|| = ||\widehat{V}_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}|| \leq ||\widehat{V}_{\mathcal{C}}^{-1/2}||||\widehat{V}_{\mathcal{C}}^{-1/2}\widehat{r}_{j,\mathcal{C}}|| \leq 2\sqrt{\frac{\tau_{\max}}{\tau_{\min}}},$$

and

$$||\boldsymbol{\eta}_2|| = ||V_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}|| \leq ||V_{\mathcal{C}}^{-1}||||\widehat{V}_{\mathcal{C}}^{1/2}||||\widehat{V}_{\mathcal{C}}^{-1/2}\widehat{r}_{j,\mathcal{C}}|| \leq \frac{2\tau_{\max}}{\tau_{\min}}.$$

Furthermore,

$$\begin{aligned} &\left|\boldsymbol{\eta}_1^{*T}\left(\widehat{V}_{\mathcal{C}} - V_{\mathcal{C}}\right)\boldsymbol{\eta}_2^*\right| \leq \max_{j_1,j_2\in\{1,2,\ldots,p\}} |\widehat{v}_{j_1,j_2} - v_{j_1,j_2}| \sum_{j_1,j_2} |\boldsymbol{\eta}_{j_1}^*||\boldsymbol{\eta}_{j_2}^*| \\ = \;& M\left(\sum_{j_1}|\boldsymbol{\eta}_{j_1}^*|\right)\left(\sum_{j_2}|\boldsymbol{\eta}_{j_2}^*|\right) \leq Mp. \end{aligned}$$

Hence,

$$\left|\widehat{r}_{j,\mathcal{C}}^T\widehat{V}_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}} - \widehat{r}_{j,\mathcal{C}}^T V_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}\right| \leq 4\left(\frac{\tau_{\max}}{\tau_{\min}}\right)^{3/2} Mp.$$

Since $||\widehat{r}_{j,\mathcal{C}} - r_{j,\mathcal{C}}|| \leq \max_{j_1,j_2\in\{1,2,\ldots,p\}} |\widehat{v}_{j_1,j_2} - v_{j_1,j_2}|\sqrt{p} = M\sqrt{p}$, $||V_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}|| \leq \frac{2\tau_{\max}}{\tau_{\min}}$, and

$||V_{\mathcal{C}}^{-1} r_{j,\mathcal{C}}|| \leq \sqrt{\frac{\tau_{\max}}{\tau_{\min}}}$, we have

$$\begin{aligned}\left|\left(\widehat{r}_{j,\mathcal{C}} - r_{j,\mathcal{C}}\right)^T V_{\mathcal{C}}^{-1}\left(\widehat{r}_{j,\mathcal{C}} + r_{j,\mathcal{C}}\right)\right| &\leq ||\widehat{r}_{j,\mathcal{C}} - r_{j,\mathcal{C}}|| \left(||V_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}}|| + ||V_{\mathcal{C}}^{-1} r_{j,\mathcal{C}}||\right) \\ &\leq \left(\frac{2\tau_{\max}}{\tau_{\min}} + \sqrt{\frac{\tau_{\max}}{\tau_{\min}}}\right) M\sqrt{p},\end{aligned}$$

and

$$\left|\widehat{r}_{j,\mathcal{C}}^T V_{\mathcal{C}}^{-1}\widehat{r}_{j,\mathcal{C}} - r_{j,\mathcal{C}}^T V_{\mathcal{C}}^{-1} r_{j,\mathcal{C}}\right| \leq \left(\frac{2\tau_{\max}}{\tau_{\min}} + \sqrt{\frac{\tau_{\max}}{\tau_{\min}}}\right) M\sqrt{p} \leq \left(2\left(\frac{\tau_{\max}}{\tau_{\min}}\right)^{3/2} + 1\right) M\sqrt{p}.$$

Therefore,

$$\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \max_{j\in\mathcal{C}^c} \left|\widehat{\sigma}_j^2 - \sigma_j^2\right| \leq 2\left(3\left(\frac{\tau_{\max}}{\tau_{\min}}\right)^{3/2} + 1\right) Mp.$$

Since $\sigma_j^2 = v_{j,j} - r_{j,\mathcal{C}}^T V_{\mathcal{C}}^{-1} r_{j,\mathcal{C}} \geq v_{j,j} \geq \tau_{\min}$,

$$\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \max_{j\in\mathcal{C}^c} \frac{\left|\widehat{\sigma}_j^2 - \sigma_j^2\right|}{\sigma_j^2} \leq \frac{2}{\tau_{\min}}\left(3\left(\frac{\tau_{\max}}{\tau_{\min}}\right)^{3/2} + 1\right) Mp.$$

Let $L = \frac{4}{\tau_{\min}}\left(3\left(\frac{\tau_{\max}}{\tau_{\min}}\right)^{3/2} + 1\right)$. For $0 < \epsilon < 1$, we have

$$\begin{aligned}&\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \max_{j\in\mathcal{C}^c} \left|\log\widehat{\sigma}_j^2 - \log\sigma_j^2\right| > \epsilon\right) \leq \Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}} \max_{j\in\mathcal{C}^c} \frac{\left|\widehat{\sigma}_j^2 - \sigma_j^2\right|}{\sigma_j^2} > \frac{\epsilon}{2}\right) \\ \leq\ & \Pr\left(M = \max_{j_1,j_2\in\{1,2,\ldots,p\}} |\widehat{v}_{j_1,j_2} - v_{j_1,j_2}| > \frac{\epsilon}{pL}\right) \leq \sum_{j_1,j_2\in\{1,2,\ldots,p\}} \Pr\left(|\widehat{v}_{j_1,j_2} - v_{j_1,j_2}| > \frac{\epsilon}{pL}\right) \\ \leq\ & \frac{p(p+1)}{2} C_1 \exp\left(-C_2 n \frac{\epsilon^2}{p^2 L^2}\right),\end{aligned}$$

where the last inequality follows from Bernstein inequality since predictors $\mathbf{X}$ follows either a multivariate normal distribution or a finite mixture of multivariate normal distributions.

Similarly,

$$\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}\max_{j\in\mathcal{C}^c}\left|\log\left[\widehat{\sigma}_j^{(h)}\right]^2-\log\left[\sigma_j^{(h)}\right]^2\right|>\epsilon\right)\leq\frac{p(p+1)}{2}C_1\exp\left(-C_2n\frac{s_h\epsilon^2}{p^2L^2}\right).$$

Thus,

$$\begin{aligned}
&\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}\max_{j\in\mathcal{C}^c}\left|\sum_{h=1}^{H}s_h\log\left[\widehat{\sigma}_j^{(h)}\right]^2-\sum_{h=1}^{H}s_h\log\left[\sigma_j^{(h)}\right]^2\right|>\epsilon\right)\\
\leq\ &\sum_{h=1}^{H}\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}\max_{j\in\mathcal{C}^c}\left|\log\left[\widehat{\sigma}_j^{(h)}\right]^2-\log\left[\sigma_j^{(h)}\right]^2\right|>\frac{\epsilon}{s_hH}\right)\\
\leq\ &\sum_{h=1}^{H}\frac{p(p+1)}{2}C_1\exp\left(-C_2n\frac{\epsilon^2}{s_hH^2p^2L^2}\right)\leq\frac{Hp(p+1)}{2}C_1\exp\left(-C_2n\frac{\epsilon^2}{H^2p^2L^2}\right).
\end{aligned}$$

□

*Proof of Theorem 2.* We denote $R_{j|\mathcal{C}}=\log\widehat{\sigma}_j^2-\log\sigma_j^2$ and

$$\widetilde{R}_{j|\mathcal{C}}=\sum_{h=1}^{H}s_h\log\left[\widehat{\sigma}_j^{(h)}\right]^2-\sum_{h=1}^{H}s_h\log\left[\sigma_j^{(h)}\right]^2.$$

According to Lemma 3, for $0<\epsilon<1$, there exists constant $C_1$ and $C_2$ such that

$$\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}\max_{j\in\mathcal{C}^c}\left|R_{j|\mathcal{C}}\right|>\epsilon\right)\leq\frac{p(p+1)}{2}C_1\exp\left(-C_2n\frac{\epsilon^2}{p^2L^2}\right),$$

and

$$\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}\max_{j\in\mathcal{C}^c}\left|\widetilde{R}_{j|\mathcal{C}}\right|>\epsilon\right)\leq\frac{Hp(p+1)}{2}C_1\exp\left(-C_2n\frac{\epsilon^2}{H^2p^2L^2}\right),$$

where $L = \frac{4}{\tau_{\min}}\left(3\left(\frac{\tau_{\max}}{\tau_{\min}}\right)^{3/2}+1\right)$. Under Condition 2, $p = o\left(n^{\rho}\right)$ and $2\rho + 2\kappa < 1$,

$$\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}\max_{j\in\mathcal{C}^c}\left|R_{j|\mathcal{C}}\right| > Cn^{-\kappa}\right) \leq \frac{p(p+1)}{2}C_1\exp\left(-C_2 n^{1-2\kappa-2\rho}\frac{C^2}{L^2}\right) \to 0,$$

$$\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}\max_{j\in\mathcal{C}^c}\left|\widetilde{R}_{j|\mathcal{C}}\right| > Cn^{-\kappa}\right) \leq \frac{Hp(p+1)}{2}C_1\exp\left(-C_2 n^{1-2\kappa-2\rho}\frac{C^2}{H^2L^2}\right) \to 0,$$

for any positive constant $C$ as $n \to \infty$. According to Proposition 5,

$$\begin{aligned}
\widehat{D}^*_{j|\mathcal{C}} &= \log\widehat{\sigma}^2_j - \sum_{h=1}^{H} s_h \log\left[\widehat{\sigma}^{(h)}_j\right]^2 \\
&= \log\sigma^2_j - \sum_{h=1}^{H} s_h \log\left[\sigma^{(h)}_j\right]^2 + R_{j|\mathcal{C}} - \widetilde{R}_{j|\mathcal{C}} \\
&= \log\left(1 + \frac{\mathrm{Var}\left(M_j\right) - \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)^T}{\mathbb{E}\left(V_j\right)}\right) + \\
&\quad \log\left(\mathbb{E}V_j\right) - \mathbb{E}\log\left(V_j\right) + R_{j|\mathcal{C}} - \widetilde{R}_{j|\mathcal{C}},
\end{aligned}$$

where $M_j = \mathbb{E}\left(X_j|\mathbf{X}_{\mathcal{C}}, S(Y)\right)$ and $V_j = \mathrm{Var}\left(X_j|\mathbf{X}_{\mathcal{C}}, S(Y)\right)$.

When $\mathcal{C}^c \cap \mathcal{A} \neq \emptyset$ and all the relevant predictors indexed by $\mathcal{A}$ are stepwise detectable with constant $\kappa \geq 0$, then there exists $m \geq 0$ such that $\cup_{i=0}^{m-1}\mathcal{T}_i \subset \mathcal{C}$ and $\mathcal{C}^c \cap \mathcal{T}_m \neq \emptyset$. According to Definition 3, there exists $j \in \mathcal{C}^c \cap \mathcal{T}_m$ and $\xi_1, \xi_2 > 0$ such that either

$$\frac{\mathrm{Var}\left(M_j\right) - \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)^T}{\mathbb{E}\left(V_j\right)} \geq \xi_1 n^{-\kappa},$$

that is, with sufficiently large $n$,

$$\log\left(1 + \frac{\mathrm{Var}\left(M_j\right) - \mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(M_j, \mathbf{X}_{\mathcal{C}}\right)^T}{\mathbb{E}\left(V_j\right)}\right) \geq \frac{\xi_1}{2}n^{-\kappa}.$$

or

$$\log\left(\mathbb{E}V_j\right) - \mathbb{E}\log\left(V_j\right) \geq \xi_2 n^{-\kappa}$$

Let $c = \min\left(\frac{\xi_1}{4}, \frac{\xi_2}{2}\right)$. Therefore,

$$\begin{aligned}
\widehat{D}^*_{j|\mathcal{C}} &\geq \log\left(1 + \frac{\mathrm{Var}\left(M_j\right) - \mathrm{Cov}\left(M_j, \mathbf{X}_\mathcal{C}\right)\left[\mathrm{Cov}\left(\mathbf{X}_\mathcal{C}\right)\right]^{-1}\mathrm{Cov}\left(M_j, \mathbf{X}_\mathcal{C}\right)^T}{\mathbb{E}\left(V_j\right)}\right) \\
&\quad + \log\left(\mathbb{E}V_j\right) - \mathbb{E}\log\left(V_j\right) - \left(\left|R_{j|\mathcal{C}}\right| + \left|\widetilde{R}_{j|\mathcal{C}}\right|\right) \\
&\geq 2cn^{-\kappa} - \left(\left|R_{j|\mathcal{C}}\right| + \left|\widetilde{R}_{j|\mathcal{C}}\right|\right)
\end{aligned}$$

Since

$$\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}\max_{j\in\mathcal{C}^c}\left|R_{j|\mathcal{C}}\right| > \frac{c}{2}n^{-\kappa}\right) \to 0,$$

and

$$\Pr\left(\max_{\mathcal{C}\subset\{1,2,\ldots,p\}}\max_{j\in\mathcal{C}^c}\left|\widetilde{R}_{j|\mathcal{C}}\right| > \frac{c}{2}n^{-\kappa}\right) \to 0,$$

we have

$$\Pr\left(\min_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{A}\neq\emptyset}\max_{j\in\mathcal{C}^c\cap\mathcal{A}}\widehat{D}^*_{j|\mathcal{C}} \geq cn^{-\kappa}\right) \to 1,$$

as $n \to \infty$.

When $\mathcal{C}^c \cap \mathcal{A} = \emptyset$ under model (2.8), for any $j \in \mathcal{C}^c$, $M_j = \mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}, S(Y)\right) = \mathbb{E}\left(X_j|\mathbf{X}_\mathcal{C}\right)$, which is a linear combination of predictors in $\mathbf{X}_\mathcal{C}$, and $V_j = \mathrm{Var}\left(X_j|\mathbf{X}_\mathcal{C}, S(Y)\right) = \mathrm{Var}\left(X_j|\mathbf{X}_\mathcal{C}\right)$, which is a constant that does not depend on $\mathbf{X}_\mathcal{C}$ or $S(Y)$. Then,

$$\frac{\mathrm{Var}\left(M_j\right) - \mathrm{Cov}\left(M_j, \mathbf{X}_\mathcal{C}\right)\left[\mathrm{Cov}\left(\mathbf{X}_\mathcal{C}\right)\right]^{-1}\mathrm{Cov}\left(M_j, \mathbf{X}_\mathcal{C}\right)^T}{\mathbb{E}\left(V_j\right)} = 0,$$

and

$$\log\left(\mathbb{E}V_j\right) - \mathbb{E}\log\left(V_j\right) = 0.$$

Thus,

$$\widehat{D}^*_{j|\mathcal{C}} \leq \left|R_{j|\mathcal{C}}\right| + \left|\widetilde{R}_{j|\mathcal{C}}\right|,$$

and

$$\Pr\left(\max_{\mathcal{C}:\mathcal{C}^c\cap\mathcal{A}=\emptyset}\max_{j\in\mathcal{C}^c}\widehat{D}^*_{j|\mathcal{C}} < Cn^{-\kappa}\right) \to 1,$$

for any positive constant $C$ as $n \to \infty$. □

## A.6 Proof of Theorem 3 in Section 2.1.3

*Proof of Theorem 3.* (a) We denote

$$\begin{aligned} D^*_j &= \log\sigma^2_j - \sum_{h=1}^{H} s_h \log\left[\sigma^{(h)}_j\right]^2 \\ &= \frac{\mathrm{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right)}{\mathbb{E}\left(\mathrm{Var}\left(X_j|S(Y)\right)\right)} + \log\mathbb{E}\left(\mathrm{Var}\left(X_j|S(Y)\right)\right) - \mathbb{E}\log\left[\mathrm{Var}\left(X_j|S(Y)\right)\right]. \end{aligned}$$

First, we will prove that

$$P\left(\max_{j\in\{1,2,\ldots,p\}}\left|\widehat{D}^*_j - D^*_j\right| > Cn^{-\kappa}\right) \to 0,$$

for any constant $C > 0$ as $n \to \infty$. There exists $C_1, C_2 > 0$ such that

$$\begin{aligned} & P\left(\max_{j\in\{1,2,\ldots,p\}} \left|\log \widehat{\sigma}_j^2 - \log \sigma_j^2\right| > \epsilon\right) \leq P\left(\max_{j\in\{1,2,\ldots,p\}} \frac{\left|\widehat{\sigma}_j^2 - \sigma_j^2\right|}{\sigma_j^2} > \frac{\epsilon}{2}\right) \\ \leq \;& P\left(\max_{j\in\{1,2,\ldots,p\}} \left|\widehat{\sigma}_j^2 - \sigma_j^2\right| > \frac{\epsilon\tau_{\min}}{2}\right) \leq \sum_{j\in\{1,2,\ldots,p\}} P\left(\left|\widehat{\sigma}_j^2 - \sigma_j^2\right| > \frac{\epsilon\tau_{\min}}{2}\right) \\ \leq \;& pC_1 \exp\left(-C_2 n \frac{\epsilon^2\tau_{\min}}{4}\right), \end{aligned}$$

where the last inequality follows from Bernstein inequality since predictors $\mathbf{X}$ follows either a multivariate normal distribution or a finite mixture of multivariate normal distributions. Since $\log(p) = o(n^\gamma)$ with $\gamma > 0$ and $\gamma + 2\kappa < 1$,

$$\begin{aligned} & P\left(\max_{j\in\{1,2,\ldots,p\}} \left|\log \widehat{\sigma}_j^2 - \log \sigma_j^2\right| > Cn^{-\kappa}\right) \leq pC_1 \exp\left(-C_2 n^{1-2\kappa} \frac{C^2\tau_{\min}}{4}\right) \\ \leq \;& C_1 \exp\left(-C_2 C^2 n^{1-2\kappa} \frac{\tau_{\min}}{4} + n^\gamma\right) \to 0, \end{aligned}$$

for any constant $C > 0$ as $n \to \infty$. Similarly,

$$\begin{aligned} & P\left(\max_{j\in\{1,2,\ldots,p\}} \left|\sum_{h=1}^{H} s_h \left(\log\left[\widehat{\sigma}_j^{(h)}\right]^2 - \log\left[\sigma_j^{(h)}\right]^2\right)\right| > Cn^{-\kappa}\right) \\ \leq \;& \sum_{h=1}^{H} P\left(\max_{j\in\{1,2,\ldots,p\}} \left|\log\left[\widehat{\sigma}_j^{(h)}\right]^2 - \log\left[\sigma_j^{(h)}\right]^2\right| > \frac{Cn^{-\kappa}}{s_h H}\right) \\ \leq \;& HpC_1 \exp\left(-C_2 n^{1-2\kappa} \frac{C^2\tau_{\min}}{4H^2}\right) \\ \leq \;& HC_1 \exp\left(-C_2 C^2 n^{1-2\kappa} \frac{\tau_{\min}}{4H^2} + n^\gamma\right) \to 0, \end{aligned}$$

for any constant $C > 0$ as $n \to \infty$. Thus,

$$\begin{aligned} & P\left(\max_{j\in\{1,2,\ldots,p\}} \left||\widehat{D}_j^* - D_j^*\right| > Cn^{-\kappa}\right) \\ = \; & P\left(\max_{j\in\{1,2,\ldots,p\}} \left|\left(\log\widehat{\sigma}_j^2 - \log\sigma_j^2\right) + \sum_{h=1}^{H} s_h\left(\log\left[\widehat{\sigma}_j^{(h)}\right]^2 - \log\left[\sigma_j^{(h)}\right]^2\right)\right| > Cn^{-\kappa}\right) \\ \to \; & 0, \end{aligned}$$

for any constant $C > 0$ as $n \to \infty$.

Second, note that if $\frac{\mathrm{Var}(\mathbb{E}(X_j|S(Y)))}{\mathbb{E}(\mathrm{Var}(X_j|S(Y)))} \geq \xi_1 n^{-\kappa}$, then for sufficiently large $n$,

$$n\log\left[1 + \frac{\mathrm{Var}\left(\mathbb{E}(X_j|S(Y))\right)}{\mathbb{E}\left(\mathrm{Var}(X_j|S(Y))\right)}\right] \geq \frac{\xi_1}{2}n^{1-\kappa}.$$

When all the relevant predictors indexed by $\mathcal{A}$ are independently detectable, there exists $c = \min\{\frac{\xi_1}{4}, \frac{\xi_2}{2}\} > 0$ such that

$$D_j^* \geq 2cn^{-\kappa}, \text{ for } j \in \mathcal{A}.$$

The events $\left\{\min_{j\in\mathcal{A}} D_j^* < cn^{-\kappa}\right\} \subset \left\{\max_{j\in\{1,2,\ldots,p\}} \left|\widehat{D}_j^* - D_j^*\right| > cn^{-\kappa}\right\}$, and

$$P\left(\min_{j\in\mathcal{A}} D_j^* < cn^{-\kappa}\right) \leq P\left(\max_{j\in\{1,2,\ldots,p\}} \left|\widehat{D}_j^* - D_j^*\right| > cn^{-\kappa}\right) \to 0,$$

as $n \to \infty$.

(b) We denote $\widehat{M_c} = \{j : \widehat{D}_j^* \geq cn^{-\kappa}, \text{ for } 1 \leq j \leq p\}$ and $M_{\frac{c}{2}} = \{j : D_j^* \geq \frac{c}{2}n^{-\kappa}, \text{ for } 1 \leq$

$j \leq p\}$. We will prove that there exists $C > 0$ such that $\left|M_{\frac{c}{2}}\right| \leq Cn^{\kappa+\eta}$. Since

$$\left|M_{\frac{c}{2}}\right| \frac{c}{2} n^{-\kappa} \leq \sum_{j=1}^{p} \widehat{D}_j^*,$$

we just need to prove that

$$\sum_{j=1}^{p} \widehat{D}_j^* \leq Cn^{\eta},$$

for some positive constant $C$. First,

$$\sum_{j=1}^{p} \log\left[1 + \frac{\mathrm{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right)}{\mathbb{E}\left(\mathrm{Var}\left(X_j|S(Y)\right)\right)}\right] \leq \sum_{j=1}^{p} \frac{\mathrm{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right)}{\mathbb{E}\left(\mathrm{Var}\left(X_j|S(Y)\right)\right)} \leq \frac{1}{\tau_{\min}} \sum_{j=1}^{p} \mathrm{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right),$$

and

$$\sum_{j=1}^{p} \mathrm{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right) = \sum_{j\in\mathcal{A}^c} \mathrm{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right) + \sum_{j\in\mathcal{A}} \mathrm{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right).$$

Under model (2.8), $\mathbb{E}\left(\mathbf{X}_{\mathcal{A}^c}|S(Y)\right) = \alpha + \boldsymbol{\beta}^T \mathbb{E}\left(\mathbf{X}_{\mathcal{A}}|S(Y)\right)$. Since

$$\begin{aligned}
&\sum_{j\in\mathcal{A}^c} \mathrm{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right) = \mathrm{trace}\left(\mathrm{Cov}\left[\mathbb{E}\left(\mathbf{X}_{\mathcal{A}^c}|S(Y)\right)\right]\right) \\
= \;& \mathrm{trace}\left(\boldsymbol{\beta}^T \mathrm{Cov}\left[\mathbb{E}\left(\mathbf{X}_{\mathcal{A}}|S(Y)\right)\right]\boldsymbol{\beta}\right) \\
\leq \;& \lambda_{\max}\left(\boldsymbol{\beta}^T\boldsymbol{\beta}\right) \mathrm{trace}\left(\mathrm{Cov}\left[\mathbb{E}\left(\mathbf{X}_{\mathcal{A}}|S(Y)\right)\right]\right) \\
= \;& \lambda_{\max}\left(\boldsymbol{\beta}^T\boldsymbol{\beta}\right) \sum_{j\in\mathcal{A}} \mathrm{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right).
\end{aligned}$$

and

$$\begin{aligned}
& \lambda_{\max}\left(\boldsymbol{\beta}^T\boldsymbol{\beta}\right) \\
\leq\ & \frac{\lambda_{\max}\left(\mathrm{Cov}\left(\mathbf{X}_{\mathcal{A}^c},\mathbf{X}_{\mathcal{A}}|S(Y)\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{A}}|S(Y)\right)\right]^{-1}\mathrm{Cov}\left(\mathbf{X}_{\mathcal{A}^c},\mathbf{X}_{\mathcal{A}}|S(Y)\right)^T\right)}{\lambda_{\min}\left(\mathrm{Cov}\left(\mathbf{X}_{\mathcal{A}}|S(Y)\right)\right)} \\
\leq\ & \frac{\lambda_{\max}\left(\mathrm{Cov}\left(\mathbf{X}_{\mathcal{A}^c}|S(Y)\right)\right)}{\lambda_{\min}\left(\mathrm{Cov}\left(\mathbf{X}_{\mathcal{A}}|S(Y)\right)\right)} \leq \frac{\tau_{\max}}{\tau_{\min}},
\end{aligned}$$

we have

$$\begin{aligned}
\sum_{j=1}^{p}\log\left[1+\frac{\mathrm{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right)}{\mathbb{E}\left(\mathrm{Var}\left(X_j|S(Y)\right)\right)}\right] &\leq \frac{1}{\tau_{\min}}\sum_{j=1}^{p}\mathrm{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right) \\
&\leq \frac{1}{\tau_{\min}}\left(1+\frac{\tau_{\max}}{\tau_{\min}}\right)\sum_{j\in\mathcal{A}}\mathrm{Var}\left(\mathbb{E}\left(X_j|S(Y)\right)\right).
\end{aligned}$$

Second, for $j\in\mathcal{A}^c$, $\mathrm{Var}(X_j|\mathbf{X}_{\mathcal{A}},S(Y))$ is a constant, and

$$\begin{aligned}
& \log\left[\mathbb{E}\left(\mathrm{Var}(X_j|S(Y))\right)\right]-\mathbb{E}\left[\log\left(\mathrm{Var}(X_j|S(Y))\right)\right] \\
\leq\ & \mathbb{E}\left(\mathrm{Var}(X_j|S(Y))\right)\mathbb{E}\left(\frac{1}{\mathrm{Var}(X_j|S(Y))}\right)-1 \leq \frac{\mathbb{E}\left(\mathrm{Var}(X_j|S(Y))\right)}{\mathrm{Var}(X_j|\mathbf{X}_{\mathcal{A}},S(Y))}-1 \\
=\ & \frac{\mathbb{E}\left[\mathrm{Var}\left(\mathbb{E}(X_j|\mathbf{X}_{\mathcal{A}},S(Y))|S(Y)\right)\right]}{\mathrm{Var}(X_j|\mathbf{X}_{\mathcal{A}},S(Y))} \leq \frac{1}{\tau_{\min}}\mathbb{E}\left[\mathrm{Var}\left(\mathbb{E}(X_j|\mathbf{X}_{\mathcal{A}},S(Y))|S(Y)\right)\right].
\end{aligned}$$

So

$$\begin{aligned}
&\sum_{j\in\mathcal{A}^c}\left(\log\left[\mathbb{E}\left(\mathrm{Var}(X_j|S(Y))\right)\right]-\mathbb{E}\left[\log\left(\mathrm{Var}(X_j|S(Y))\right)\right]\right)\\
\leq\ & \frac{1}{\tau_{\min}}\sum_{j\in\mathcal{A}^c}\mathbb{E}\left[\mathrm{Var}\left(\mathbb{E}(X_j|\mathbf{X}_{\mathcal{A}},S(Y))|S(Y)\right)\right]\\
=\ & \frac{1}{\tau_{\min}}\mathrm{trace}\left(\mathbb{E}\left[\mathrm{Cov}\left(\mathbb{E}(\mathbf{X}_{\mathcal{A}^c}|\mathbf{X}_{\mathcal{A}},S(Y))|S(Y)\right)\right]\right)=\frac{1}{\tau_{\min}}\mathrm{trace}\left(\boldsymbol{\beta}^T\mathbb{E}\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{A}}|S(Y)\right)\right]\boldsymbol{\beta}^T\right)\\
\leq\ & \frac{1}{\tau_{\min}}\lambda_{\max}\left(\boldsymbol{\beta}^T\boldsymbol{\beta}\right)\sum_{j\in\mathcal{A}}\mathbb{E}\left[\mathrm{Var}\left(X_j|S(Y)\right)\right]\leq\frac{\tau_{\max}}{\tau_{\min}^2}\sum_{j\in\mathcal{A}}\mathbb{E}\left[\mathrm{Var}\left(X_j|S(Y)\right)\right].
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{j=1}^{p}\left(\log\left[\mathbb{E}\left(\mathrm{Var}(X_j|S(Y))\right)\right]-\mathbb{E}\left[\log\left(\mathrm{Var}(X_j|S(Y))\right)\right]\right)\\
\leq\ & \frac{1}{\tau_{\min}}\left(1+\frac{\tau_{\max}}{\tau_{\min}}\right)\sum_{j\in\mathcal{A}}\mathbb{E}\left[\mathrm{Var}\left(X_j|S(Y)\right)\right].
\end{aligned}$$

Moreover, $\sum_{j\in\mathcal{A}}\mathrm{Var}\left(X_j|S(Y)\right)\leq|\mathcal{A}|\tau_{\max}\leq\tau_{\max}\xi_0 n^{\eta}$, and

$$\begin{aligned}
&\sum_{j=1}^{p}D_j^*\\
=\ & \sum_{j=1}^{p}\left(\log\left[1+\frac{\mathrm{Var}\left(\mathbb{E}(X_j|S(Y))\right)}{\mathbb{E}\left(\mathrm{Var}(X_j|S(Y))\right)}\right]+\log\left[\mathbb{E}\left(\mathrm{Var}(X_j|S(Y))\right)\right]-\mathbb{E}\left[\log\left(\mathrm{Var}(X_j|S(Y))\right)\right]\right)\\
\leq\ & \frac{1}{\tau_{\min}}\left(1+\frac{\tau_{\max}}{\tau_{\min}}\right)\sum_{j\in\mathcal{A}}\mathrm{Var}\left(X_j|S(Y)\right)\leq\frac{\tau_{\max}}{\tau_{\min}}\left(1+\frac{\tau_{\max}}{\tau_{\min}}\right)\xi_0 n^{\eta}.
\end{aligned}$$

Thus, there exists $C>0$ such that

$$\left|M_{\frac{c}{2}}\right|\leq\frac{2}{c}n^{\kappa}\sum_{j=1}^{p}D_j^*\leq\frac{\tau_{\max}}{\tau_{\min}}\left(1+\frac{\tau_{\max}}{\tau_{\min}}\right)\frac{2\xi_0}{c}n^{\kappa+\eta}\leq Cn^{\kappa+\eta}.$$

Then,

$$P\left(\left|\widehat{M}_c\right| > Cn^{\kappa+\eta}\right) \leq P\left(\left|\widehat{M}_c\right| > \left|M_{\frac{c}{2}}\right|\right),$$

and the event $\left\{\left|\widehat{M}_c\right| > \left|M_{\frac{c}{2}}\right|\right\} \subset \left\{\text{there exists } j \text{ such that } D_j^* < \frac{c}{2}n^{-\kappa} \text{ and } \widehat{D}_j^* \geq cn^{-\kappa}\right\} \subset \left\{\max_{j\in\{1,2,\ldots,p\}} \left|\widehat{D}_j^* - D_j^*\right| > \frac{c}{2}n^{-\kappa}\right\}$. Thus, according to the results in (a),

$$P\left(\left|\widehat{M}_c\right| > Cn^{\kappa+\eta}\right) \leq P\left(\max_{j\in\{1,2,\ldots,p\}} \left|\widehat{D}_j^* - D_j^*\right| > \frac{c}{2}n^{-\kappa}\right) \to 0,$$

as $n \to \infty$. □

## A.7 Proof of Choices of Slicing Schemes in Section 2.2

Suppose $(S_1, S_2, \ldots, S_H)$ is the true slicing scheme under model (2.3) or (2.8), and $S(Y)$ denotes the slice membership of the slicing scheme, *i.e.*, $S(Y) = h$ if $Y \in S_h$. We say that a slicing scheme $\widetilde{S}(Y)$ is a refinement of $S(Y)$, which is denoted by $\widetilde{S}(Y) \preceq S(Y)$, if there exists a function $g$ such that $S(Y) = g(\widetilde{S}(Y))$.

For any slicing scheme $\widetilde{S}(Y)$, as $n \to \infty$, the limit of the likelihood-ratio test statistic under model (2.3) is given by

$$\begin{aligned}
& D_{j|\mathcal{C},\widetilde{S}(Y)} = \lim_{n\to\infty} \widehat{D}_{j|\mathcal{C},\widetilde{S}(Y)} \\
= \;& \log\left[\mathrm{Var}\left(X_j\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C}\right)\left[\mathrm{Var}\left(\mathbf{X}_\mathcal{C}\right)\right]^{-1}\mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C}\right)^T\right] - \\
& \log\left[\mathbb{E}\left(\mathrm{Var}\left(X_j \mid \widetilde{S}(Y)\right)\right) - \right. \\
& \left.\mathbb{E}\left(\mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)\right)\left[\mathbb{E}\left(\mathrm{Cov}\left(\mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)\right)\right]^{-1}\mathbb{E}\left(\mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)\right)^T\right],
\end{aligned}$$

and the limit of the augmented likelihood-ratio test statistic under model (2.8) is given by

$$
\begin{aligned}
& D^*_{j|\mathcal{C},\widetilde{S}(Y)} = \lim_{n\to\infty} \widehat{D}^*_{j|\mathcal{C},\widetilde{S}(Y)} \\
= \;& \log\left[\mathrm{Var}\left(X_j\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Var}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)^T\right] - \\
& \mathbb{E}\left(\log\left[\mathrm{Var}\left(X_j \mid \widetilde{S}(Y)\right) - \right.\right. \\
& \left.\left.\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right]^{-1}\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)^T\right]\right)
\end{aligned}
$$

For the true slicing scheme $S(Y)$ or a slicing scheme $\widetilde{S}(Y)$ that is a refinement of $S(Y)$, *i.e.*, $\widetilde{S}(Y) \preceq S$, under model (2.3), we have

$$
\begin{aligned}
& D_{j|\mathcal{C},\widetilde{S}(Y)} = D_{j|\mathcal{C},S(Y)} \\
= \;& \log\left[\mathrm{Var}\left(X_j\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)^T\right] \\
& - \log\left(\mathrm{Var}(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y))\right),
\end{aligned}
$$

where $\mathrm{Var}(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y))$ is a constant that does not depend on $\mathbf{X}_{\mathcal{C}}$ or $S(Y)$. Similary, under the augmented model (2.8),

$$
\begin{aligned}
& D^*_{j|\mathcal{C},\widetilde{S}(Y)} = D^*_{j|\mathcal{C},S(Y)} \\
= \;& \log\left[\mathrm{Var}\left(X_j\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)^T\right] \\
& -\mathbb{E}\left(\log\left(\mathrm{Var}(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y))\right)\right).
\end{aligned}
$$

For a slicing scheme $\widetilde{S}(Y)$ that is "coarser" than the true slicing scheme $S(Y)$, *i.e.*, $S(Y) \preceq \widetilde{S}(Y)$, we have the following theorem.

**Proposition 6.** *Suppose $\widetilde{S}(Y)$ is a slicing scheme such that $S(Y) \preceq \widetilde{S}(Y)$, where $S(Y)$ is*

*the true slicing scheme.*

*(a) Under model (2.3),*

$$D_{j|\mathcal{C},S(Y)} \geq D_{j|\mathcal{C},\widetilde{S}(Y)},$$

*where equality holds if $\mathcal{A} \subset \mathcal{C}$, where $\mathcal{A}$ is the index set of relevant predictors.*

*(b) Under model (2.8),*

$$D^*_{j|\mathcal{C},S(Y)} \geq D^*_{j|\mathcal{C},\widetilde{S}(Y)},$$

*where equality holds if $\mathcal{A} \subset \mathcal{C}$, where $\mathcal{A}$ is the index set of relevant predictors.*

*Proof of Proposition 6.* (a) Since

$$\begin{aligned}\mathrm{Var}\left(X_j \mid \widetilde{S}(Y)\right) &= \mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right) \mid \widetilde{S}(Y)\right) \\ &\quad + \mathrm{Var}\left(\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right) \mid \widetilde{S}(Y)\right),\end{aligned}$$

we have

$$\begin{aligned}&\mathbb{E}\left(\mathrm{Var}\left(X_j \mid \widetilde{S}(Y)\right)\right) - \\ &\mathbb{E}\left(\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right)\left[\mathbb{E}\left(\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right)\right]^{-1}\mathbb{E}\left(\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right)^T \\ =\ &\mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right)\right) + \mathbb{E}\left[\mathrm{Var}\left(\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right) \mid \widetilde{S}(Y)\right)\right] - \\ &\mathbb{E}\left[\mathrm{Cov}\left(\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right), \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right]\left[\mathbb{E}\left(\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right)\right]^{-1} \\ &\mathbb{E}\left[\mathrm{Cov}\left(\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right), \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right]^T \\ \geq\ &\mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right)\right),\end{aligned}$$

where equality holds if $\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right)$ is a linear combination of $\mathbf{X}_{\mathcal{C}}$ that does not depend

on $\widetilde{S}(Y)$. Since $S(Y) \preceq \widetilde{S}(Y)$, the $\sigma$-algebra $\sigma(\widetilde{S}(Y)) \subset \sigma(S(Y))$. Thus,

$$\begin{aligned}
\operatorname{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right) &= \mathbb{E}\left(\operatorname{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right) \middle| \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right) \\
&\quad + \operatorname{Var}\left(\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right) \middle| \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right) \\
&\geq \mathbb{E}\left(\operatorname{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right) \middle| \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right),
\end{aligned}$$

and

$$\mathbb{E}\left(\operatorname{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right)\right) \geq \mathbb{E}\left(\operatorname{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)\right),$$

where equality holds if $\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)$ is a linear combination of $\mathbf{X}_{\mathcal{C}}$ that does not depend on $S(Y)$. Because $S(Y)$ is the true slicing scheme, under model (2.3), $\operatorname{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)$ is a constant that does not depend on $\mathbf{X}_{\mathcal{C}}$ or $S(Y)$, that is,

$$\mathbb{E}\left(\operatorname{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)\right) = \operatorname{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right).$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left(\operatorname{Var}\left(X_j \mid \widetilde{S}(Y)\right)\right) - \\
&\mathbb{E}\left(\operatorname{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right)\left[\mathbb{E}\left(\operatorname{Cov}\left(\mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right)\right]^{-1}\mathbb{E}\left(\operatorname{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right)^T \\
\geq\ &\operatorname{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right),
\end{aligned}$$

$$
\begin{aligned}
& D_{j|\mathcal{C},S(Y)} \\
= \; & \log\left[\mathrm{Var}\left(X_j\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C}\right)\left[\mathrm{Cov}\left(\mathbf{X}_\mathcal{C}\right)\right]^{-1}\mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C}\right)^T\right] - \\
& \log\left(\mathrm{Var}(X_j \mid \mathbf{X}_\mathcal{C}, S(Y))\right) \\
\geq \; & \log\left[\mathrm{Var}\left(X_j\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C}\right)\left[\mathrm{Cov}\left(\mathbf{X}_\mathcal{C}\right)\right]^{-1}\mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C}\right)^T\right] - \\
& \log\left[\mathbb{E}\left(\mathrm{Var}\left(X_j \mid \widetilde{S}(Y)\right)\right) - \right. \\
& \left. \mathbb{E}\left(\mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)\right)\left[\mathbb{E}\left(\mathrm{Cov}\left(\mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)\right)\right]^{-1}\mathbb{E}\left(\mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)\right)^T\right] \\
= \; & D_{j|\mathcal{C},\widetilde{S}(Y)}.
\end{aligned}
$$

When $\mathcal{A} \subset \mathcal{C}$, under model (2.3), the predictor indexed by $j \in \mathcal{C}^c$ has the same conditional distribution across difference slices given $\mathbf{X}_\mathcal{C}$. So $\mathbb{E}\left(X_j \mid \mathbf{X}_\mathcal{C}, S(Y)\right)$ and $\mathbb{E}\left(X_j \mid \mathbf{X}_\mathcal{C}, \widetilde{S}(Y)\right)$ are linear combinations of $\mathbf{X}_\mathcal{C}$ that do not depend on $S(Y)$ or $\widetilde{S}(Y)$. Thus, the equalities hold in this case.

To prove (b), note that

$$
\begin{aligned}
& \mathrm{Var}\left(X_j \mid \widetilde{S}(Y)\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)\left[\mathrm{Cov}\left(\mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)\right]^{-1}\mathrm{Cov}\left(X_j, \mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)^T \\
= \; & \mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_\mathcal{C}, \widetilde{S}(Y)\right) \mid \widetilde{S}(Y)\right) + \mathrm{Var}\left(\mathbb{E}\left(X_j \mid \mathbf{X}_\mathcal{C}, \widetilde{S}(Y)\right) \mid \widetilde{S}(Y)\right) - \\
& \mathrm{Cov}\left(\mathbb{E}\left(X_j \mid \mathbf{X}_\mathcal{C}, \widetilde{S}(Y)\right), \mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)\left[\mathrm{Cov}\left(\mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)\right]^{-1} \\
& \mathrm{Cov}\left(\mathbb{E}\left(X_j \mid \mathbf{X}_\mathcal{C}, \widetilde{S}(Y)\right), \mathbf{X}_\mathcal{C} \mid \widetilde{S}(Y)\right)^T \\
\geq \; & \mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_\mathcal{C}, \widetilde{S}(Y)\right) \mid \widetilde{S}(Y)\right).
\end{aligned}
$$

where equality holds if $\mathbb{E}\left(X_j \mid \mathbf{X}_\mathcal{C}, \widetilde{S}(Y)\right)$ is a linear combination of $\mathbf{X}_\mathcal{C}$ that does not depend

on $\widetilde{S}(Y)$. Since $S(Y) \preceq \widetilde{S}(Y)$, the $\sigma$-algebra $\sigma(\widetilde{S}(Y)) \subset \sigma(S(Y))$. Thus,

$$\begin{aligned}
\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right) &= \mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)\middle|\mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right) \\
&\quad + \mathrm{Var}\left(\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)\middle|\mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right) \\
&\geq \mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)\middle|\mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right),
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{Var}\left(X_j \mid \widetilde{S}(Y)\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right]^{-1}\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)^T \\
\geq\ & \mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right)\middle|\widetilde{S}(Y)\right) \geq \mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)\middle|\widetilde{S}(Y)\right),
\end{aligned}$$

where the equalities hold if $\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)$ and $\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right)$ are linear combinations of $\mathbf{X}_{\mathcal{C}}$ that do not depend on $S(Y)$ or $\widetilde{S}(Y)$. According to Jensen's inequality,

$$\log\left[\mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)\middle|\widetilde{S}(Y)\right)\right] \geq \mathbb{E}\left(\log\left[\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)\right]\middle|\widetilde{S}(Y)\right).$$

where equality holds if $\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)$ is a constant that does not depend on $\mathbf{X}_{\mathcal{C}}$ or $S(Y)$.

Therefore,

$$
\begin{aligned}
& \mathbb{E}\left(\log\left[\mathrm{Var}\left(X_j \mid \widetilde{S}(Y)\right) - \right.\right. \\
& \left.\left. \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right]^{-1} \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)^T\right]\right) \\
\geq\ & \mathbb{E}\left(\log\left[\mathbb{E}\left(\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right)\middle|\widetilde{S}(Y)\right)\right]\right) \\
\geq\ & \mathbb{E}\left(\mathbb{E}\left(\log\left[\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)\right] \mid \widetilde{S}(Y)\right)\right) \\
=\ & \mathbb{E}\left(\log\left[\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)\right]\right), \\
& D^*_{j|\mathcal{C},S(Y)} \\
=\ & \log\left[\mathrm{Var}\left(X_j\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)^T\right] - \\
& \mathbb{E}\left(\log\left[\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)\right]\right) \\
\geq\ & \log\left[\mathrm{Var}\left(X_j\right) - \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}}\right)\right]^{-1}\mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}}\right)^T\right] - \\
& \mathbb{E}\left(\log\left[\mathrm{Var}\left(X_j \mid \widetilde{S}(Y)\right) - \right.\right. \\
& \left.\left. \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\left[\mathrm{Cov}\left(\mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)\right]^{-1} \mathrm{Cov}\left(X_j, \mathbf{X}_{\mathcal{C}} \mid \widetilde{S}(Y)\right)^T\right]\right) \\
=\ & D^*_{j|\mathcal{C},\widetilde{S}(Y)}.
\end{aligned}
$$

When $\mathcal{A} \subset \mathcal{C}$ under the augmented model (2.8), the predictor indexed by $j \in \mathcal{C}^c$ has the same conditional distribution across difference slices given $\mathbf{X}_{\mathcal{C}}$. So $\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)$ and $\mathbb{E}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, \widetilde{S}(Y)\right)$ are linear combinations of $\mathbf{X}_{\mathcal{C}}$ that do not depend on $S(Y)$ or $\widetilde{S}(Y)$, and $\mathrm{Var}\left(X_j \mid \mathbf{X}_{\mathcal{C}}, S(Y)\right)$ is a constant that does not depend on $\mathbf{X}_{\mathcal{C}}$ or $S(Y)$. Thus, the equalities hold in this case. □

# Appendix B

# Miscellaneous

## B.1 Forward-Summation Algorithm in Section 3.3.2

Without loss of generality, assume individuals are sorted with $R_d = \{1, 2, \ldots, N\}$. Given individual ranks $R_d$ and variances $\{\sigma_c^2 : J_c = d\}$, we can sum over all possible individual partitions under the sequential partition prior by filling entries of the table $\left[W_j^*\right]_{0 \le j \le N}$ ($W_0^* \equiv 1$) recursively,

$$W_j^* = \sum_{i=0}^{j-1} \left( \frac{\pi_T}{1-\pi_T} W_i^* U_{i+1,j}^* V_{i+1,j}^* \right), \tag{B.1}$$

with

$$U_{i+1,j}^* = \prod_{c:J_c=d} \sqrt{\frac{(n_c+\kappa_1)\kappa_2}{(n_c+\kappa_1)\kappa_2 + n_c(j-i)\kappa_1}} \exp\left( \frac{\kappa_1^2 \left( \sum_{s=i+1}^{j} \sum_{C_j=c} y_{s,j} \right)^2}{2\sigma_c^2 (n_c+\kappa_1) \left[ (n_c+\kappa_1)\kappa_2 + n_c(j-i)\kappa_1 \right]} \right),$$

and

$$V_{i+1,j}^* = \prod_{m=1}^{M_d} \frac{\Gamma(\lambda)}{\Gamma(\lambda + j - i)} \prod_{h=1}^{2^{|\mathcal{A}_{d,m}|}} \frac{\Gamma\left( n_{h,d} + \frac{\lambda}{2^{|\mathcal{A}_{d,m}|}} \right)}{\Gamma\left( \frac{\lambda}{2^{|\mathcal{A}_{d,m}|}} \right)},$$

where $\frac{\pi_T}{1-\pi_T}$ is the prior odds for adding a new individual type and $n_{h,d}$ is the number of individuals from $\{i+1,\ldots,j\}$ who have the $h$th genotype among $2^{|\mathcal{A}_{d,m}|}$ possible genotype combinations from markers in $\mathcal{A}_{d,m}$.

The joint probability of observing gene expression $\{\mathbf{Y}_{\mathcal{K}_c} : J_c = d\}$ and genotypes of genetic markers $\mathbf{X}_{\mathcal{A}_d}$ marginalizing over individual type partition $T_d$ can be obtained from the last entry of the table $\left[W_j^*\right]_{0\le j\le N}$:

$$\begin{aligned}
& \Pr\left(\mathbf{X}_{\mathcal{A}_d}, \{\mathbf{Y}_{\mathcal{K}_c} : J_c = d\} | R_d, \{\sigma_c^2 : J_c = d\}\right) \\
= \; & W_N^*(1-\pi_T)^N \prod_{c:J_c=d} \left(2\pi\sigma_c^2\right)^{\frac{Nn_c}{2}} \left(\sqrt{\frac{\kappa_1}{n_c+\kappa_1}}\right)^N \\
& \times \exp\left( -\frac{\sum_{i=1}^N \left[\sum_{C_j=c} y_{i,j}^2 - \frac{\left(\sum_{C_j=c} y_{i,j}\right)^2}{n_c+\kappa_1}\right]}{2\sigma_c^2} \right).
\end{aligned}$$

To sample individual type partitions $T_d$, or equivalently the slice boundaries, we can use backward sampling based on the table $\left[W_j^*\right]_{0\le j\le N}$.

## B.2 MCMC Algorithm and Convergence Diagnostics in Section 4.2.1

To obtain posterior distribution of parameters and hidden indicators according to the ANOVA model (4.1, we can iterate the following Markov Chain Monte Carlo (MCMC) sampling algorithm:

1. Given $\tau_j$, $\omega_f^{\pm}$, $\delta_i^{\pm}$, $\sigma^2$ and $\gamma_2$, iteratively sample $Q_{i,j}$ (with $F_i = f$) by first marginalizing over $\omega_{f,j}$ (integrating out $\omega_{f,j}$ given $P_{f,j}$) and then summing over $P_{f,j} \in \{0,1,2\}$;

2. Given $\tau_j$, $\omega_f^{\pm}$, $\delta_i^{\pm}$, $\sigma^2$ and $\gamma_2$, iteratively sample $P_{f,j}$ by marginalizing over $\omega_{f,j}$ and conditioning on $Q_{i,j}$ (with $F_i = f$), and then sample $\omega_{f,j}$ conditioning on $P_{f,j}$;

3. Given $\tau_j$, $\omega_{f,j}$ and $\sigma^2$, sample $\delta_i^{\pm}$ conditioning on $Q_{i,j}$ (with $F_i = f$) and obtain $\delta_{i,j}$ directly;

4. Given $\omega_{f,j}$, $\sigma^2$ and $\gamma_2$, sample $\omega_f^{\pm}$ conditioning on $P_{f,j}$;

5. Given $\omega_f^{\pm}$, $\delta_i^{\pm}$, $\gamma_1$ and $\gamma_2$, sample $\sigma^2$ by integrating out $\tau_j$ and $\omega_{f,j}$ and conditioning on $P_{f,j}$ and $Q_{i,j}$ (the marginal posterior of $\sigma^2$ is similar to the form given in Section 4.3.1);

6. Given $\omega_{f,j}$, $\delta_{i,j}$, $\sigma^2$ and $\gamma_1$, sample $\tau_j$;

7. Given $\tau_j$ and $\sigma^2$, sample $\gamma_1$;

8. Given $\omega_{f,j}$, $\omega_f^{\pm}$ and $\sigma^2$, sample $\gamma_2$ conditioning on $P_{f,j}$.

The convergence diagnostics of the MCMC algorithm are given in Figure B.1. The Gelman and Rubin's potential scale reduction factors (Gelman and Rubin, 1992) for posterior draws based on five independent MCMC runs (after independent burn-in periods) are 1.007 for log-likelihood, 1.007 for the variance parameter $\sigma^2$, 1.003 for the background noise scale parameter $\gamma_1$ and 1.026 for the family-wise effect scale parameter $\gamma_2$.

## B.3 MCMC Algorithm, Convergence Diagnostics and Sensitivity Analysis in Section 4.3.1

We can draw subclass assignments $\mathbf{C} = \{C_i : 1 \leq i \leq N_T\}$, group assignment $\mathbf{G} = \{G_j : 1 \leq j \leq N_K\}$ and indicators of preferred groups $\mathbf{I} = \{I_g : 1 \leq g \leq N_G\}$ from the posterior distribution $\Pr(\mathbf{C}, \mathbf{G}, \mathbf{I}|\mathbf{Y})$ by iterating the following Gibbs sampler steps:

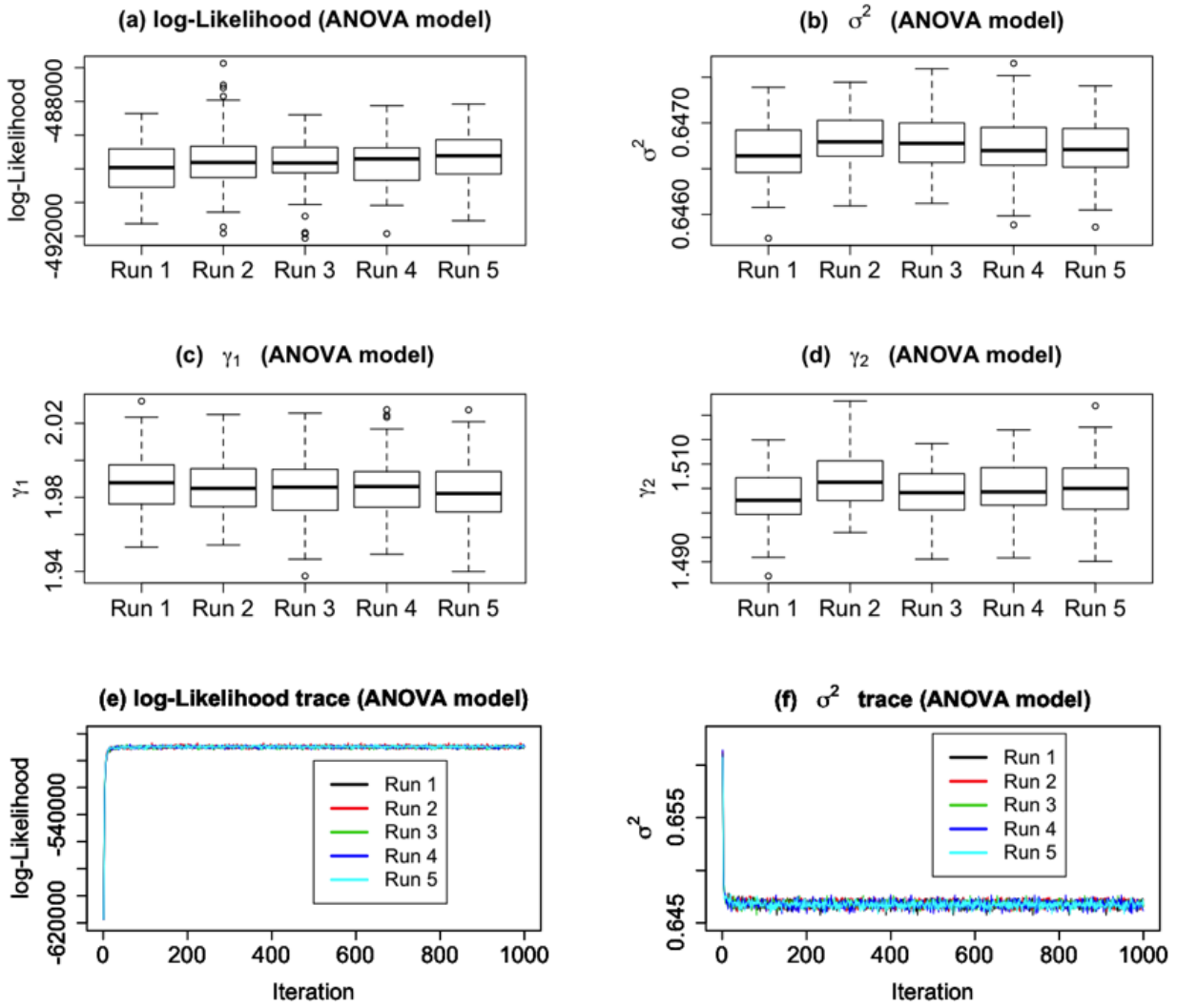


Figure B.1: MCMC diagnostics for ANOVA model. Posterior draws based on five independent runs showing boxplots of (a) log-likelihood (with a Gelman and Rubin's potential scale reduction factor 1.007); (b) variance parameter $\sigma^2$ (with a Gelman and Rubin's potential scale reduction factor 1.007); (c) background noise scale parameter $\gamma_1$ (with a Gelman and Rubin's potential scale reduction factor 1.003); (d) family-wise effect scale parameter $\gamma_2$ (with a Gelman and Rubin's potential scale reduction factor 1.026); and trace plots of (e) log-likelihood and (f) variance $\sigma^2$ from five MCMC runs.

1. Sample group assignment $\mathbf{G} = \{G_j : 1 \leq j \leq N_K\}$ conditioning on $\mathbf{C}$ and $\mathbf{I}$;

2. Sample indicators of preferred groups $\mathbf{I} = \{I_g : 1 \leq g \leq N_G\}$ conditioning on $\mathbf{C}$ and $\mathbf{G}$;

3. Iteratively sample subclass assignment of the $i$th TF, $C_i$, conditioning $\mathbf{I}$, $\mathbf{G}$ and $\mathbf{C}_{[-i]} = \{C_k : k \neq i\}$, according to the Chinese Restaurant Process (CRP) representation of the Dirichlet process.

The convergence diagnostics of the MCMC algorithm are given in Figure B.2. The Gelman and Rubin's potential scale reduction factor (Gelman and Rubin, 1992) for posterior draws of log-likelihood based on five independent MCMC runs (after the burn-in period) is 1.008.

To assess the influence of choices of hyper-parameters on the inferences drawn, we conducted sensitivity analysis on the total number of groups $N_G$, variance and scale parameters $\sigma_0^2$, $\kappa_1$ and $\kappa_2$. The inference procedure based on Bayesian partition model was repeated under the following 11 combinations of hyper-parameter settings:

1. $N_G = 100$, $\kappa_1 = 1.0$, $\kappa_2 = 1.0$ and $\sigma_0^2 = 1.0$ (priors used in obtaining the results in Section 4.4);

2. $N_G = 100$, $\kappa_1 = 1.0$, $\kappa_2 = 1.0$ and $\sigma_0^2 = 0.5$;

3. $N_G = 100$, $\kappa_1 = 1.0$, $\kappa_2 = 0.5$ and $\sigma_0^2 = 1.0$;

4. $N_G = 100$, $\kappa_1 = 0.5$, $\kappa_2 = 1.0$ and $\sigma_0^2 = 1.0$;

5. $N_G = 100$, $\kappa_1 = 1.0$, $\kappa_2 = 1.0$ and $\sigma_0^2 = 2.0$;

6. $N_G = 100$, $\kappa_1 = 1.0$, $\kappa_2 = 2.0$ and $\sigma_0^2 = 1.0$;

7. $N_G = 100$, $\kappa_1 = 2.0$, $\kappa_2 = 1.0$ and $\sigma_0^2 = 1.0$;

8. $N_G = 100$, $\kappa_1 = 0.5$, $\kappa_2 = 0.5$ and $\sigma_0^2 = 0.5$;

9. $N_G = 100$, $\kappa_1 = 2.0$, $\kappa_2 = 2.0$ and $\sigma_0^2 = 2.0$;

10. $N_G = 50$, $\kappa_1 = 1.0$, $\kappa_2 = 1.0$ and $\sigma_0^2 = 1.0$;

11. $N_G = 200$, $\kappa_1 = 1.0$, $\kappa_2 = 1.0$ and $\sigma_0^2 = 1.0$;

The primary quantities of interest, posterior probabilities on subclass preferred binding for $k$-mers, were calculated under each of the above hyper-parameter combination. Figure B.3 shows the pair-wise correlations of posterior probabilities on subclass preferred binding for each pair of hyper-parameter settings. Given a fixed number of groups, the inferences on posterior probabilities on subclass preferred binding are highly correlated (with correlations above 0.9) under different hyper-parameters. When we compare two very different specifications on the number of groups, $N_G = 50$ versus $N_G = 200$, the correlation is 0.65. When we compare $N_G = 100$ (which is used in obtaining the results in Section 4.4) with $N_G = 50$ or $N_G = 200$, all the correlations are above 0.8 even under different combinations of other hyper-parameters.

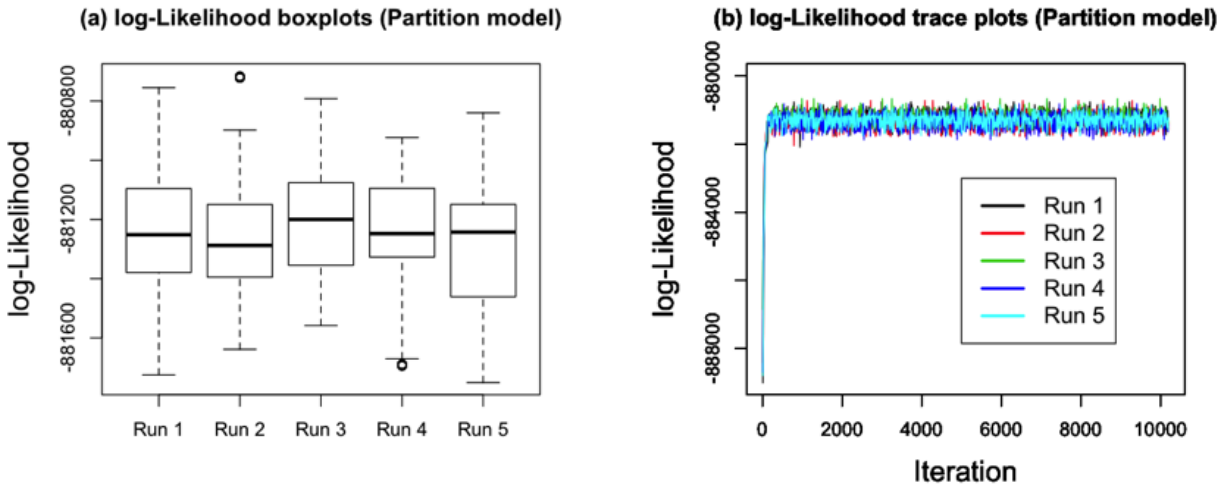


Figure B.2: MCMC convergence diagnostics and prior sensitivity analysis of Bayesian partition model in Section 2.3.1. using ETS DBD class data. Posterior draws based on five independent runs showing (a) boxplots and (b) trace plots of log-likelihood. The Gelman and Rubin's potential scale reduction factor is 1.008.

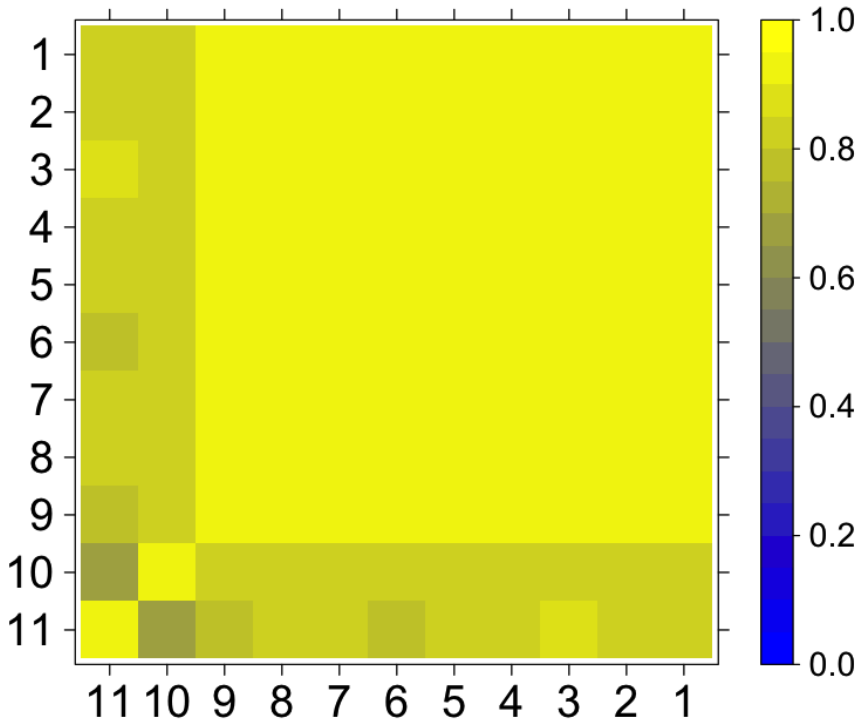


Figure B.3: Correlations of posterior probabilities on subclass preferred binding for $k$-mers under different priors. See Appendix B.3 for prior settings 1-11.

# Bibliography

Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., Metzler, G., Vedenko, A., Chen, X., et al. (2009), "Diversity and complexity in DNA recognition by transcription factors," *Science*, 324, 1720–1723.

Badis, G., Chan, E. T., van Bakel, H., Pena-Castillo, L., Tillo, D., Tsui, K., Carlson, C. D., Gossett, A. J., Hasinoff, M. J., Warren, C. L., et al. (2008), "A library of yeast transcription factor motifs reveals a widespread function for Rsc3 in targeting nucleosome exclusion at promoters," *Molecular Cell*, 32, 878–887.

Beer, M. A. and Tavazoie, S. (2004), "Predicting gene expression from sequence," *Cell*, 117, 185–198.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society: Series B (Methodological)*, 289–300.

Benos, P. V., Lapedes, A. S., and Stormo, G. D. (2002), "Is there a code for protein–DNA recognition? Probab (ilistical) ly," *Bioessays*, 24, 466–475.

Berger, M. F., Badis, G., Gehrke, A. R., Talukder, S., Philippakis, A. A., Pena-Castillo, L., Alleyne, T. M., Mnaimneh, S., Botvinnik, O. B., Chan, E. T., et al. (2008), "Variation in

homeodomain DNA binding revealed by high-resolution analysis of sequence preferences," *Cell*, 133, 1266–1276.

Berger, M. F. and Bulyk, M. L. (2009), "Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors," *Nature Protocols*, 4, 393–411.

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., and Bulyk, M. L. (2006), "Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities," *Nature Biotechnology*, 24, 1429–1435.

Berger, R. L. and Hsu, J. C. (1996), "Bioequivalence trials, intersection-union tests and equivalence confidence sets," *Statistical Science*, 11, 283–319.

Bien, J., Taylor, J., and Tibshirani, R. (2012), "A Lasso for hierarchical interactions," *arXiv preprint arXiv:1205.5050*.

Bing, N. and Hoeschele, I. (2005), "Genetical genomics analysis of a yeast segregant population for transcription network inference," *Genetics*, 170, 533–542.

Brem, R. B. and Kruglyak, L. (2005), "The landscape of genetic complexity across 5,700 gene expression traits in yeast," *Proceedings of the National Academy of Sciences of the United States of America*, 102, 1572–1577.

Brem, R. B., Yvert, G., Clinton, R., and Kruglyak, L. (2002), "Genetic dissection of transcriptional regulation in budding yeast," *Science*, 296, 752–755.

Bulyk, M. L., Huang, X., Choo, Y., and Church, G. M. (2001), "Exploring the DNA-binding

specificities of zinc fingers with DNA microarrays," *Proceedings of the National Academy of Sciences*, 98, 7158–7163.

Busser, B. W., Shokri, L., Jaeger, S. A., Gisselbrecht, S. S., Singhania, A., Berger, M. F., Zhou, B., Bulyk, M. L., and Michelson, A. M. (2012), "Molecular mechanism underlying the regulatory specificity of a Drosophila homeodomain protein that specifies myoblast identity," *Development*, 139, 1164–1174.

Bystrykh, L., Weersing, E., Dontje, B., Sutton, S., Pletcher, M. T., Wiltshire, T., Su, A. I., Vellenga, E., Wang, J., Manly, K. F., et al. (2005), "Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'," *Nature Genetics*, 37, 225–232.

Campbell, T. L., De Silva, E. K., Olszewski, K. L., Elemento, O., and Llinás, M. (2010), "Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite," *PLoS Pathogens*, 6, e1001165.

Chen, C.-H. and Li, K.-C. (1998), "Can SIR be as popular as multiple linear regression?" *Statistica Sinica*, 8, 289–316.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., Jiang, J., et al. (2008a), "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells," *Cell*, 133, 1106–1117.

Chen, Y., Zhu, J., Lum, P. Y., Yang, X., Pinto, S., MacNeil, D. J., Zhang, C., Lamb, J., Edwards, S., Sieberts, S. K., et al. (2008b), "Variations in DNA elucidate molecular networks that cause disease," *Nature*, 452, 429–435.

Cheng, Y. and Church, G. M. (2000), "Biclustering of expression data," in *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, vol. 8, pp. 93–103.

Chesler, E. J., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H. C., Mountz, J. D., Baldwin, N. E., Langston, M. A., et al. (2005), "Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function," *Nature Genetics*, 37, 233–242.

Chun, H. and Keleş, S. (2009), "Expression quantitative trait loci mapping with multivariate sparse partial least squares regression," *Genetics*, 182, 79–90.

Cloonan, N., Forrest, A. R., Kolle, G., Gardiner, B. B., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S., Bethel, G., et al. (2008), "Stem cell transcriptome profiling via massive-scale mRNA sequencing," *Nature Methods*, 5, 613–619.

Cockerham, C. C. (1954), "An extension of the concept of partitioning hereditary variance for analysis of covariances among relatives when epistasis is present," *Genetics*, 39, 859.

Cook, R. D. (2007), "Fisher lecture: Dimension reduction in regression," *Statistical Science*, 22, 1–26.

De Masi, F., Grove, C. A., Vedenko, A., Alibés, A., Gisselbrecht, S. S., Serrano, L., Bulyk, M. L., and Walhout, A. J. (2011), "Using a structural and logics systems approach to infer bHLH–DNA binding specificity determinants," *Nucleic Acids Research*, 39, 4553–4563.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least angle regression," *The Annals of Statistics*, 32, 407–499.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998), "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, 95, 14863–14868.

Fakhouri, W. D., Ay, A., Sayal, R., Dresch, J., Dayringer, E., and Arnosti, D. N. (2010), "Deciphering a transcriptional regulatory code: modeling short-range repression in the Drosophila embryo," *Molecular Systems Biology*, 6.

Fan, J. and Li, R. (2001), "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J. and Lv, J. (2008), "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70, 849–911.

Fisher, R. A. (1919), "The correlation between relatives on the supposition of Mendelian inheritance." *Transactions of the Royal Society of Edinburgh*, 52, 399–433.

Fong, A. P., Yao, Z., Zhong, J. W., Cao, Y., Ruzzo, W. L., Gentleman, R. C., and Tapscott, S. J. (2012), "Genetic and epigenetic determinants of neurogenesis and myogenesis," *Developmental Cell*.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007), "Pathwise coordinate optimization," *The Annals of Applied Statistics*, 1, 302–332.

Gelman, A. and Rubin, D. B. (1992), "Inference from iterative simulation using multiple sequences," *Statistical Science*, 7, 457–472.

Geman, S. and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 721–741.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999), "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, 286, 531–537.

Gordân, R., Hartemink, A. J., and Bulyk, M. L. (2009), "Distinguishing direct versus indirect transcription factor–DNA interactions," *Genome Research*, 19, 2090–2100.

Gordân, R., Murphy, K. F., McCord, R. P., Zhu, C., Vedenko, A., Bulyk, M. L., et al. (2011), "Curated collection of yeast transcription factor DNA binding specificity data reveals novel structural and gene regulatory insights," *Genome Biology*, 12, R125.

Grove, C. A., De Masi, F., Barrasa, M. I., Newburger, D. E., Alkema, M. J., Bulyk, M. L., and Walhout, A. J. (2009), "A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors," *Cell*, 138, 314–327.

Gusenleitner, D., Howe, E. A., Bentink, S., Quackenbush, J., and Culhane, A. C. (2012), "iBBiG: iterative binary bi-clustering of gene sets," *Bioinformatics*.

Harbison, C. T., Gordon, D. B., Lee, T. I., Rinaldi, N. J., Macisaac, K. D., Danford, T. W., Hannett, N. M., Tagne, J.-B., Reynolds, D. B., Yoo, J., et al. (2004), "Transcriptional regulatory code of a eukaryotic genome," *Nature*, 431, 99–104.

Hollenhorst, P. C., Chandler, K. J., Poulsen, R. L., Johnson, W. E., Speck, N. A., and

Graves, B. J. (2009), "DNA specificity determinants associate with distinct transcription factor functions," *PLoS Genetics*, 5, e1000778.

Hubner, N., Wallace, C. A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V., et al. (2005), "Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease," *Nature Genetics*, 37, 243–253.

Jiang, C. and Zeng, Z.-B. (1995), "Multiple trait analysis of genetic mapping for quantitative trait loci," *Genetics*, 140, 1111–1127.

Johnson, W. E., Li, C., and Rabinovic, A. (2007), "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics*, 8, 118–127.

Kendziorski, C., Chen, M., Yuan, M., Lan, H., and Attie, A. (2006), "Statistical methods for expression quantitative trait loci (eQTL) mapping," *Biometrics*, 62, 19–27.

Lan, H., Chen, M., Flowers, J. B., Yandell, B. S., Stapleton, D. S., Mata, C. M., Mui, E. T.-K., Flowers, M. T., Schueler, K. L., Manly, K. F., et al. (2006), "Combined expression trait correlations and expression quantitative trait locus mapping," *PLoS Genetics*, 2, e6.

Lander, E. S. and Botstein, D. (1989), "Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps," *Genetics*, 121, 185–199.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., Wootton, J. C., et al. (1993), "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment," *Science*, 262, 208–214.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012), "The sva package for removing batch effects and other unwanted variation in high-throughput experiments," *Bioinformatics*, 28, 882–883.

Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., Geman, D., Baggerly, K., and Irizarry, R. A. (2010), "Tackling the widespread and critical impact of batch effects in high-throughput data," *Nature Reviews Genetics*, 11, 733–739.

Leek, J. T. and Storey, J. D. (2007), "Capturing heterogeneity in gene expression studies by surrogate variable analysis," *PLoS Genetics*, 3, e161.

— (2008), "A general framework for multiple testing dependence," *Proceedings of the National Academy of Sciences*, 105, 18718–18723.

Li, H., Lu, L., Manly, K. F., Chesler, E. J., Bao, L., Wang, J., Zhou, M., Williams, R. W., and Cui, Y. (2005), "Inferring gene transcriptional modulatory relations: a genetical genomics approach," *Human Molecular Genetics*, 14, 1119–1125.

Li, K.-C. (1991), "Sliced inverse regression for dimension reduction," *Journal of the American Statistical Association*, 86, 316–327.

Li, L. (2007), "Sparse sufficient dimension reduction," *Biometrika*, 94, 603–613.

Li, R., Zhong, W., and Zhu, L. (2012), "Feature screening via distance correlation learning," *arXiv preprint arXiv:1205.4701*.

Liu, J. S. (1994), "The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem," *Journal of the American Statistical Association*, 89, 958–966.

Luscombe, N. M., Austin, S. E., Berman, H. M., and Thornton, J. M. (2000), "An overview of the structures of protein-DNA complexes," *Genome Biology*, 1, reviews001.

Mangin, B., Thoquet, P., and Grimsley, N. (1998), "Pleiotropic QTL analysis," *Biometrics*, 54, 88–99.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, 21, 1087.

Miller, A. J. (1984), "Selection of subsets of regression variables," *Journal of the Royal Statistical Society: Series A (General)*, 389–425.

Morley, M., Molony, C. M., Weber, T. M., Devlin, J. L., Ewens, K. G., Spielman, R. S., and Cheung, V. G. (2004), "Genetic analysis of genome-wide variation in human gene expression," *Nature*, 430, 743–747.

Murphy, T. B., Dean, N., and Raftery, A. E. (2010), "Variable selection and updating in model-based discriminant analysis for high dimensional data with food authenticity applications," *The Annals of Applied Statistics*, 4, 396.

Noyes, M. B., Christensen, R. G., Wakabayashi, A., Stormo, G. D., Brodsky, M. H., and Wolfe, S. A. (2008), "Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites," *Cell*, 133, 1277–1289.

Ouyang, Z., Zhou, Q., and Wong, W. H. (2009), "ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells," *Proceedings of the National Academy of Sciences*, 106, 21521–21526.

Philippakis, A. A., Qureshi, A. M., Berger, M. F., and Bulyk, M. L. (2008), "Design of compact, universal DNA microarrays for protein binding microarray experiments," *Journal of Computational Biology*, 15, 655–665.

Robasky, K. and Bulyk, M. L. (2011), "UniPROBE, update 2011: expanded content and search tools in the online database of protein-binding microarray data on protein–DNA interactions," *Nucleic Acids Research*, 39, D124–D128.

Royston, P. and Sauerbrei, W. (2008), "Interactions between treatment and continuous covariates: a step toward individualizing therapy," *Journal of Clinical Oncology*, 26, 1397–1399.

Rubin, D. B. (1978), "Bayesian inference for causal effects: The role of randomization," *The Annals of Statistics*, 34–58.

— (2005), "Causal inference using potential outcomes," *Journal of the American Statistical Association*, 100, 322–331.

Schadt, E. E., Lamb, J., Yang, X., Zhu, J., Edwards, S., GuhaThakurta, D., Sieberts, S. K., Monks, S., Reitman, M., Zhang, C., et al. (2005), "An integrative genomics approach to infer causal associations between gene expression and disease," *Nature Genetics*, 37, 710–717.

Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., et al. (2008), "Mapping the genetic architecture of gene expression in human liver," *PLoS Biology*, 6, e107.

Schadt, E. E., Monks, S. A., Drake, T. A., Lusis, A. J., Che, N., Colinayo, V., Ruff, T. G.,

Milligan, S. B., Lamb, J. R., Cavet, G., et al. (2003), "Genetics of gene expression surveyed in maize, mouse and man," *Nature*, 422, 297–302.

Schneider, T. D. and Stephens, R. M. (1990), "Sequence logos: a new way to display consensus sequences," *Nucleic Acids Research*, 18, 6097–6100.

Senger, K., Armstrong, G. W., Rowell, W. J., Kwan, J. M., Markstein, M., and Levine, M. (2004), "Immunity regulatory DNAs share common organizational features in Drosophila," *Molecular Cell*, 13, 19–32.

Simon, N. and Tibshirani, R. (2012), "A permutation approach to testing interactions in many dimensions," *arXiv preprint arXiv:1206.6519*.

Storey, J. D., Akey, J. M., and Kruglyak, L. (2005), "Multiple locus linkage analysis of genomewide expression in yeast," *PLoS Biology*, 3, e267.

Storey, J. D. and Tibshirani, R. (2003), "Statistical significance for genomewide studies," *Proceedings of the National Academy of Sciences*, 100, 9440–9445.

Suzuki, M. and Yagi, N. (1994), "DNA recognition code of transcription factors in the helix-turn-helix, probe helix, hormone receptor, and zinc finger families," *Proceedings of the National Academy of Sciences*, 91, 12357–12361.

Szretter, M. E. and Yohai, V. J. (2009), "The sliced inverse regression algorithm as a maximum likelihood procedure," *Journal of Statistical Planning and Inference*, 139, 3570–3578.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, 267–288.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002), "Diagnosis of multiple cancer types by shrunken centroids of gene expression," *Proceedings of the National Academy of Sciences*, 99, 6567–6572.

Tiwari, H. K. and Elston, R. C. (1997), "Deriving components of genetic variance for multilocus models," *Genetic Epidemiology*, 14, 1131–1136.

Viterbi, A. (1967), "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *Information Theory, IEEE Transactions on*, 13, 260–269.

Wang, H. (2009), "Forward regression for ultra-high dimensional variable screening," *Journal of the American Statistical Association*, 104, 1512–1524.

Warner, J. B., Philippakis, A. A., Jaeger, S. A., He, F. S., Lin, J., and Bulyk, M. L. (2008), "Systematic identification of mammalian regulatory motifs' target genes and functions," *Nature Methods*, 5, 347–353.

Wei, G.-H., Badis, G., Berger, M. F., Kivioja, T., Palin, K., Enge, M., Bonke, M., Jolma, A., Varjosalo, M., Gehrke, A. R., et al. (2010), "Genome-wide analysis of ETS-family DNA-binding in vitro and in vivo," *The EMBO Journal*, 29, 2147–2160.

Yvert, G., Brem, R. B., Whittle, J., Akey, J. M., Foss, E., Smith, E. N., Mackelprang, R., Kruglyak, L., et al. (2003), "Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors," *Nature Genetics*, 35, 57–64.

Zhang, W., Zhu, J., Schadt, E. E., and Liu, J. S. (2010), "A Bayesian partition method for detecting pleiotropic and epistatic eQTL modules," *PLoS Computational Biology*, 6, e1000642.

Zhang, Y. (2012), "A novel bayesian graphical model for genome-wide multi-SNP association mapping," *Genetic Epidemiology*.

Zhang, Y. and Liu, J. S. (2007), "Bayesian inference of epistatic interactions in case-control studies," *Nature Genetics*, 39, 1167–1173.

Zhong, W., Zeng, P., Ma, P., Liu, J. S., and Zhu, Y. (2005), "RSIR: regularized sliced inverse regression for motif discovery," *Bioinformatics*, 21, 4169–4175.

Zhong, W., Zhang, T., Zhu, Y., and Liu, J. S. (2012), "Correlation pursuit: forward stepwise variable selection for index models," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Zhu, C., Byers, K. J., McCord, R. P., Shi, Z., Berger, M. F., Newburger, D. E., Saulrieta, K., Smith, Z., Shah, M. V., Radhakrishnan, M., et al. (2009), "High-resolution DNA-binding specificity analysis of yeast transcription factors," *Genome Research*, 19, 556–566.

Zhu, J., Lum, P., Lamb, J., GuhaThakurta, D., Edwards, S., Thieringer, R., Berger, J., Wu, M., Thompson, J., Sachs, A., et al. (2004), "An integrative genomics approach to the reconstruction of gene networks in segregating populations," *Cytogenetic and Genome Research*, 105, 363–374.

Zhu, J., Zhang, B., Smith, E. N., Drees, B., Brem, R. B., Kruglyak, L., Bumgarner, R. E., and Schadt, E. E. (2008), "Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks," *Nature Genetics*, 40, 854–861.

Zou, H. (2006), "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.